Why Propensity Scores Should Be Used for Matching

Ben Jann

University of Bern, ben.jann@soz.unibe.ch

Social Science Research Colloquium
TU-Kaiserslautern, June 21, 2017

Contents

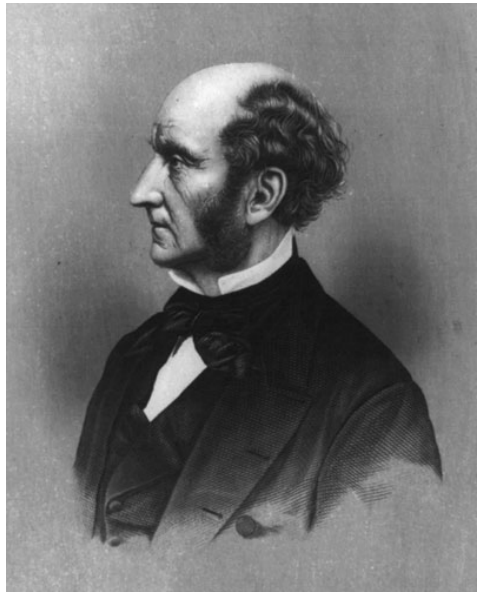
1. Potential Outcomes and Causal Inference
2. Matching
3. Propensity Score Matching
4. King and Nielsen's "Why Propensity Scores Should Not Be Used for Matching"
5. Are King and Nielsen right?
6. Illustration using `kmatch`
7. Conclusions

Counterfactual Causality (see Neyman 1923, Rubin 1974, 1990)

a.k.a. Rubin Causal Model a.k.a. Potential Outcomes Framework

- John Stuart Mill (1806–1873)

> *Thus, if a person eats of a particular dish, and dies in consequence, that is, would not have died if he had not eaten of it, people would be apt to say that eating of that dish was the cause of his death.* (Mill 2002[1843]:214)



Counterfactual Causality (see Neyman 1923, Rubin 1974, 1990)

a.k.a. Rubin Causal Model a.k.a. Potential Outcomes Framework

- Treatment variable D

$$D = \begin{cases} 1 & \text{treatment (eats from particular dish)} \\ 0 & \text{control (does not eat from particular dish)} \end{cases}$$

- Potential outcomes Y^1 and Y^0
  - Y^1: potential outcome with treatment ($D = 1$)
    - ★ If person i would eat from the particular dish, would she die or would she survive?
  - Y^0: potential outcome without treatment ($D = 0$)
    - ★ If person i would *not* eat from the particular dish, would she die or would she survive?
- Causal effect of the treatment for individual i:

causal effect = difference between potential outcomes

$$\delta_i = Y_i^1 - Y_i^0$$

Fundamental Problem of Causal Inference

- The causal effect of D on Y for individual i is defined as the difference in potential outcomes: $\delta_i = Y_i^1 - Y_i^0$
- However, the observed outcome variable is

$$Y_i = \begin{cases} Y_i^1 & \text{if } D_i = 1 \\ Y_i^0 & \text{if } D_i = 0 \end{cases}$$

- That is, only one of the two potential outcomes will be realized and, hence, only Y_i^1 or Y_i^0 can be observed, but never both.
- Consequence:

The individual treatment effect δ_i cannot be observed!

Approaches to Solve the Problem (see, e.g., Holland 1986)

- How are conclusions about causal effects be drawn despite the fundamental problem that individual treatment effects are unobservable?

- "Scientific" solution
  - Use assumptions about stability of potential outcomes over time or homogeneity of potential outcomes between units to identify causal effects.
  - Lab experiments in the natural sciences are often based on this approach.

- "Statistical" solution
  - Estimation of average causal effects based on statistical comparison of groups.

The Statistical Solution

- Instead of trying to determine individual causal effects, the statistical solution focusses on the *average* causal effect in a population (the so-called "Average Treatment Effect")
- The average causal effect can be identified by comparing the expected values of Y^1 and Y^0 since

$$ATE = E[\delta] = E[Y^1 - Y^0] = E[Y^1] - E[Y^0]$$

- Other quantities of interest:
  - Average Treatment Effect on the Treated (ATT)

$$ATT = E[Y^1 - Y^0|D = 1] = E[Y^1|D = 1] - E[Y^0|D = 1]$$

  - Average Treatment Effect on the Untreated (ATC)

$$ATT = E[Y^1 - Y^0|D = 0] = E[Y^1|D = 0] - E[Y^0|D = 0]$$

The Statistical Solution

- To determine the average effect, unbiased estimates of $E[Y^0]$ and $E[Y^1]$ are required.
- If the independence assumption

$$(Y^0, Y^1) \perp\!\!\!\perp D$$

applies, that is, if D is independent from (or at least uncorrelated with) Y^0 and Y^1, then $E[Y^0] = E[Y^0|D = 0]$ and $E[Y^1] = E[Y^1|D = 1]$.

- In this case the average causal effect can be be measured by a simple group comparison (mean difference) of observations without treatment ($D = 0$) and observations with treatment ($D = 1$).
- Randomized experiments solve the problem: If D is randomly assigned, it is independent from Y^0 and Y^1 by design.

1 Potential Outcomes and Causal Inference

2 Matching

3 Propensity Score Matching

4 King and Nielsen's "Why Propensity Scores Should Not Be Used for Matching"

5 Are King and Nielsen right?

6 Illustration using `kmatch`

7 Conclusions

Conditional Independence / Strong Ignorability

- Can causal effects also be identified from "observational" (i.e. non-experimental) data?
- Sometimes it can be argued that the independence assumption is valid *conditionally* (conditional independence, "unconfoundedness"):

$$(Y^0, Y^1) \perp\!\!\!\perp D \,|\, X$$

- If, in addition, the overlap assumption

$$0 < \Pr(D = 1|X = x) < 1, \quad \text{for all } x$$

  is given, then the ATE (or ATT or ATC) can be identified by conditioning on X.
- For example:

$$ATE = \sum_x \Pr[X = x] \{E[Y|D = 1, X = x] - E[Y|D = 0, X = x]\}$$

Estimation Under Strong Ignorability: Matching

- Basic idea
  - For each observation, find matching observations from the other group with the same (or at least very similar) X values.
  - The Y values of these matching observations are then used to compute the counterfactual outcome for the observation at hand.
  - An estimate for the average causal effect is given as the mean of the differences between the observed values and the "imputed" counterfactual values over all observations.

Estimation Under Strong Ignorability: Matching

- Formally:

$$\widehat{ATT} = \frac{1}{N^{D=1}} \sum_{i|D=1} \left[Y_i - \hat{Y}_i^0 \right] = \frac{1}{N^{D=1}} \sum_{i|D=1} \left[Y_i - \sum_{j|D=0} w_{ij} Y_j \right]$$

$$\widehat{ATC} = \frac{1}{N^{D=0}} \sum_{i|D=0} \left[\hat{Y}_i^1 - Y_i \right] = \frac{1}{N^{D=0}} \sum_{i|D=0} \left[\sum_{j|D=1} w_{ij} Y_j - Y_i \right]$$

$$\widehat{ATE} = \frac{N^{D=1}}{N} \cdot \widehat{ATT} + \frac{N^{D=0}}{N} \cdot \widehat{ATC}$$

- Different matching algorithms use different definitions of w_{ij}.

Exact Matching

- Exact matching:

$$w_{ij} = \begin{cases} 1/k_i & \text{if } X_i = X_j \\ 0 & \text{else} \end{cases}$$

  with k_i as the number of observations for which $X_i = X_j$ applies.
- The result equivalent to "perfect stratification" or "subclassification" (see, e.g., Cochran 1968).
- Problem: If X contains several variables there is a large probability that no exact matches can be found (the "curse of dimensionality").
- Coarsened Exact Matching (Blackwell et al. 2009)
  - Like exact matching, but the variables in X are "coarsened" beforehand to reduce the number of possible combinations of values (e.g. classification of continuous variables).

Mahalanobis Matching

- An alternative is to match based on a distance metric that measures the proximity between observations in the multivariate space of X.
- A common approach is to use

$$MD(X_i, X_j) = \sqrt{(X_i - X_j)' \Sigma^{-1} (X_i - X_j)}$$

  as distance metric, where Σ is an appropriate scaling matrix.
- Mahalanobis matching: Σ is the covariance matrix of X.
- Euclidean matching: Σ is the identity matrix
- Mahalanobis matching is equivalent to Euclidean matching based on standardized and orthogonalized X.

Mahalanobis Matching

- Various matching algorithms can then be employed to find potential matches (observations that are close in terms of MD) and determine the matching weights w_{ij}.
- Pair matching (one-to-one matching without replacement)
  - For each observation i in one group find observation j in the other group that has the smallest MD_{ij}. Once observation j is used as a match, do not use it again.
- Nearest-neighbor matching
  - For each observation i in one group find the k closest observations in the other group. A single observation can be used multiple times as a match. In case of ties (i.e. identical MD), use all ties as matches.
- Caliper matching
  - Like nearest-neighbor matching, but only use observations for which MD is smaller than some value c.

Mahalanobis Matching

- Radius matching
  - Use all observations as matches for which MD is smaller than some value c.
- Kernel matching
  - Like radius matching, but give larger weight to observations with smaller MD using some kernel function (such as, e.g., the Epanechnikov kernel).
- Furthermore, since matching is no longer exact, it may make sense to refine the estimates by applying regression-adjustment to the matched data (known as "bias-adjustment" in case of nearest-neighbor matching).

1 Potential Outcomes and Causal Inference

2 Matching

3 Propensity Score Matching

4 King and Nielsen's "Why Propensity Scores Should Not Be Used for Matching"

5 Are King and Nielsen right?

6 Illustration using `kmatch`

7 Conclusions

The Propensity Score Theorem (Rosenbaum and Rubin 1983)

- If the conditional independence assumption is true, then

$$\Pr(D_i = 1 | Y_i^0, Y_i^1, X_i) = \Pr(D_i = 1 | X_i) = \pi(X_i)$$

  where $\pi(X)$ is called the propensity score.

- That is,

$$(Y^0, Y^1) \perp\!\!\!\perp D \,|\, X$$

  implies

$$(Y^0, Y^1) \perp\!\!\!\perp D \,|\, \pi(X)$$

  so that under strong ignorability the average causal effect can be estimated by conditioning on the propensity score $\pi(X)$ instead of X.

- This is remarkable, because the information in X, which may include many variables, can be reduced to just one dimension. This greatly simplifies the matching task.

Propensity Score Matching (PSM)

- Instead of computing multivariate distances, we can thus simply match on the (one-dimensional) propensity score.
- Procedure
  - Step 1: Estimate the propensity score, e.g. using a Logit model.
  - Step 2: Apply a matching algorithm as above, but use differences in the propensity score, $|\hat{\pi}(X_i) - \hat{\pi}(X_j)|$, instead of the multivariate distances MD_{ij}.
- PSM is tremendously popular
  - https://scholar.google.ch/scholar?q="propensity+score"+AND+(matching+OR+matched+OR+match)

1 Potential Outcomes and Causal Inference

2 Matching

3 Propensity Score Matching

4 King and Nielsen's "Why Propensity Scores Should Not Be Used for Matching"

5 Are King and Nielsen right?

6 Illustration using `kmatch`

7 Conclusions

King and Nielsen

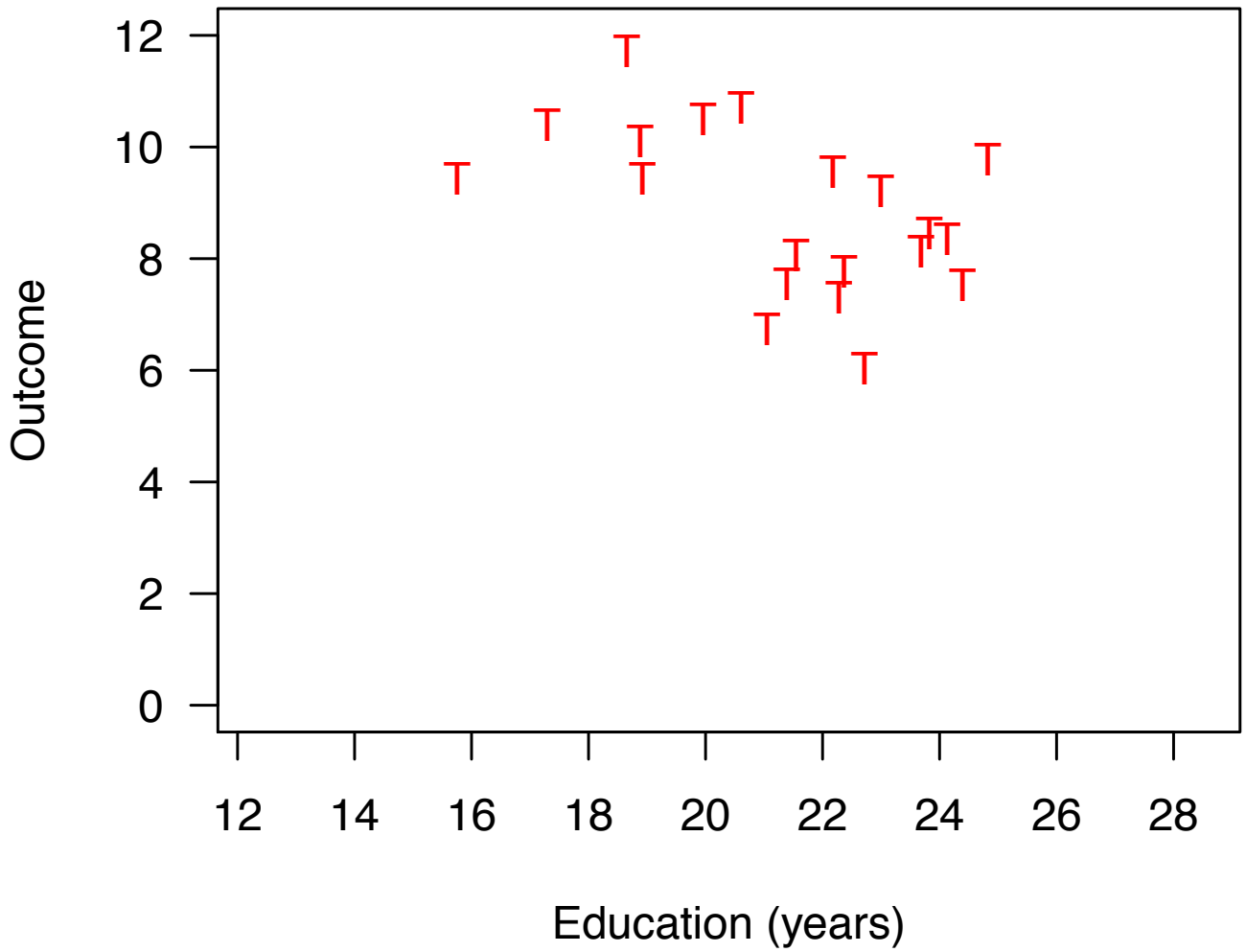
- In 2015/2016 Gary King and Richard Nielsen circulated a paper that created quite some confusion among applied researchers.
- The basic message of the paper is that PSM is really, really bad and should best be discarded.
- The paper
  - http://j.mp/1sexgVw
- Slides
  - https://gking.harvard.edu/presentations/why-propensity-scores-should-not-be-used-matching-6
- Watch it
  - https://www.youtube.com/watch?v=rBv39pK1iEs

King and Nielsen

- The story goes about as follows.
- Argument 1
  - Model dependence (i.e. dependence of results on modeling decisions made by the researcher) is bad because it leads to bias (people are selective in their choices even if they try not to be).
  - Matching is good because it reduces model dependence.

Matching to Reduce Model Dependence

(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)

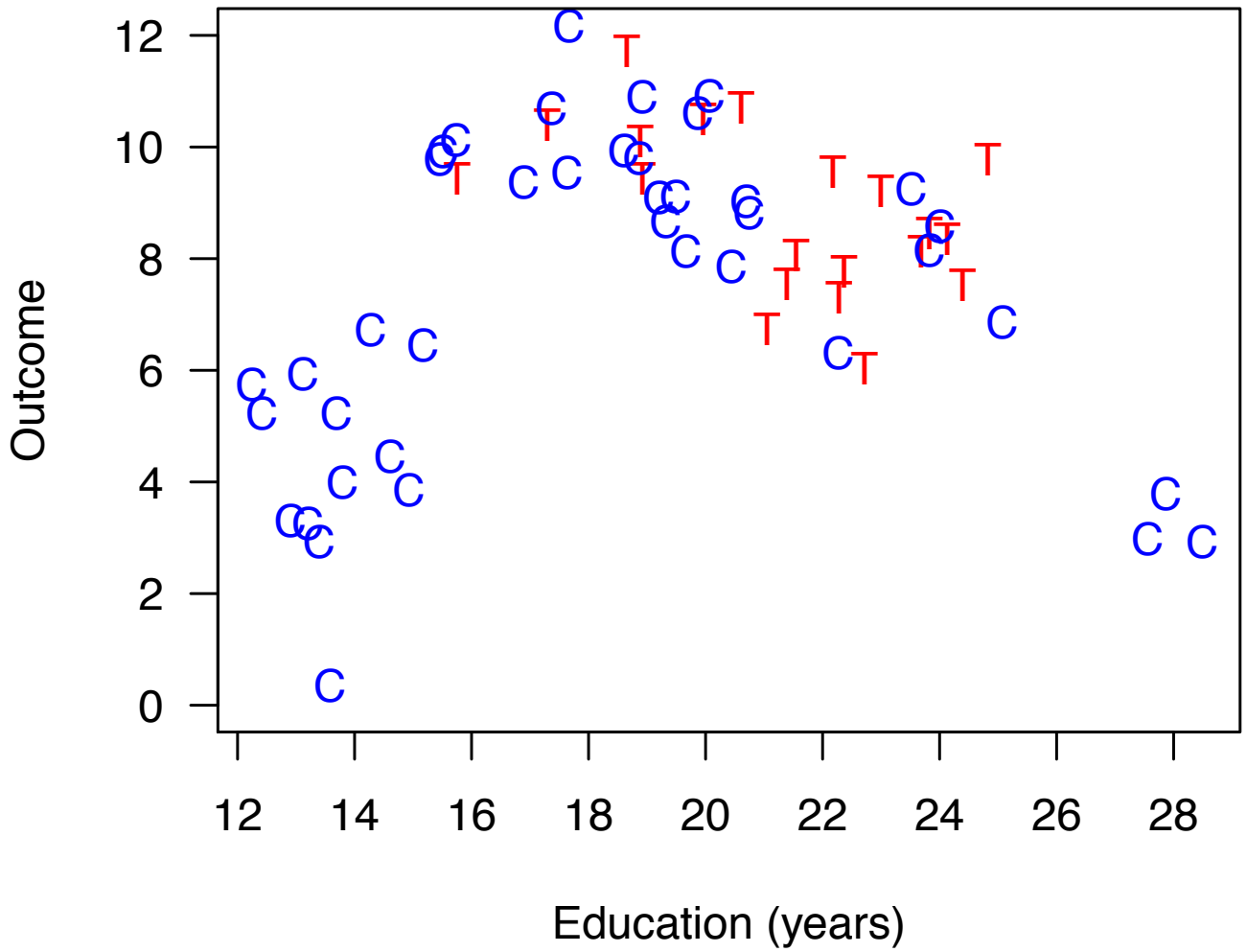


3/23

(slides by King and Nielsen)

Matching to Reduce Model Dependence

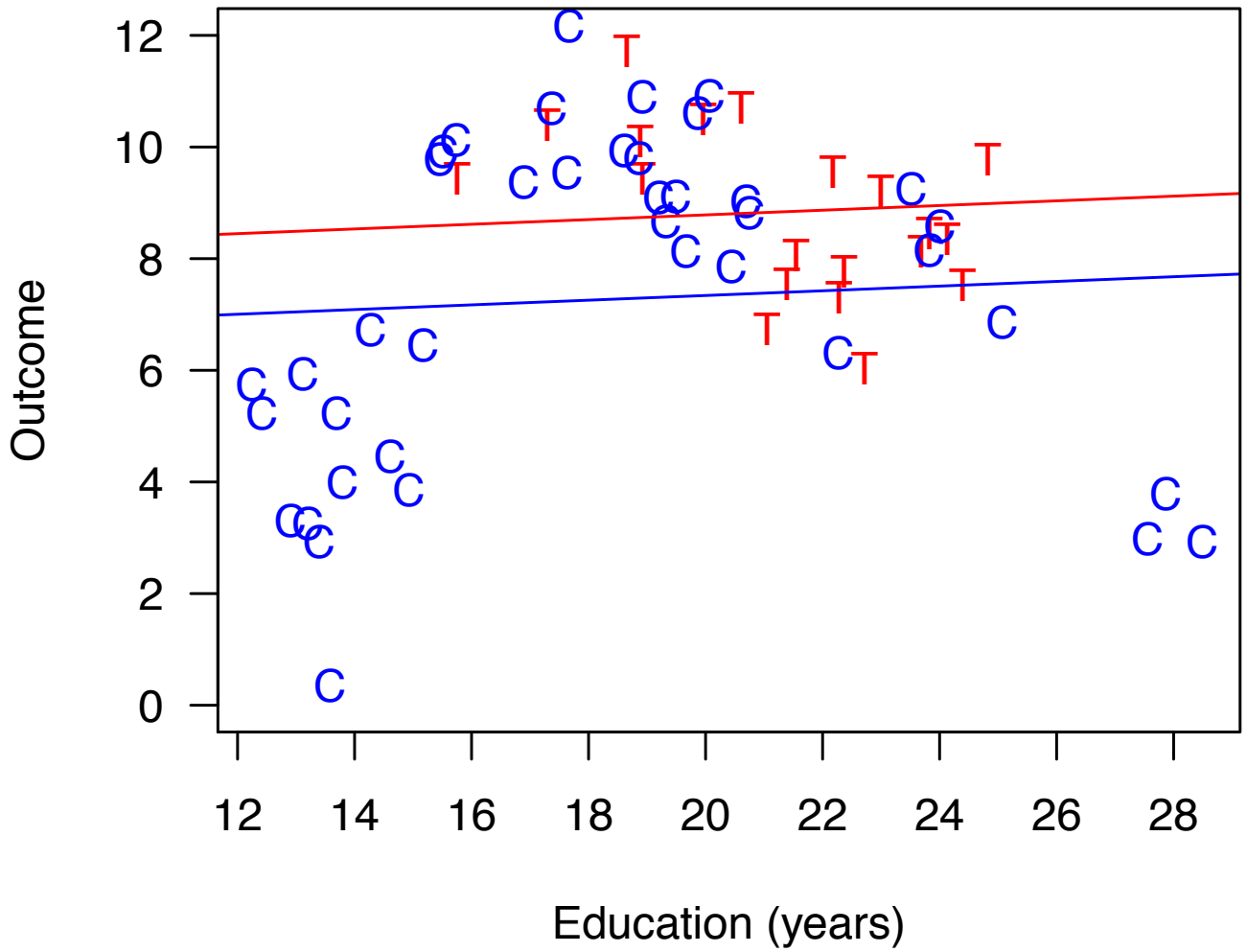
(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)



(slides by King and Nielsen)

Matching to Reduce Model Dependence

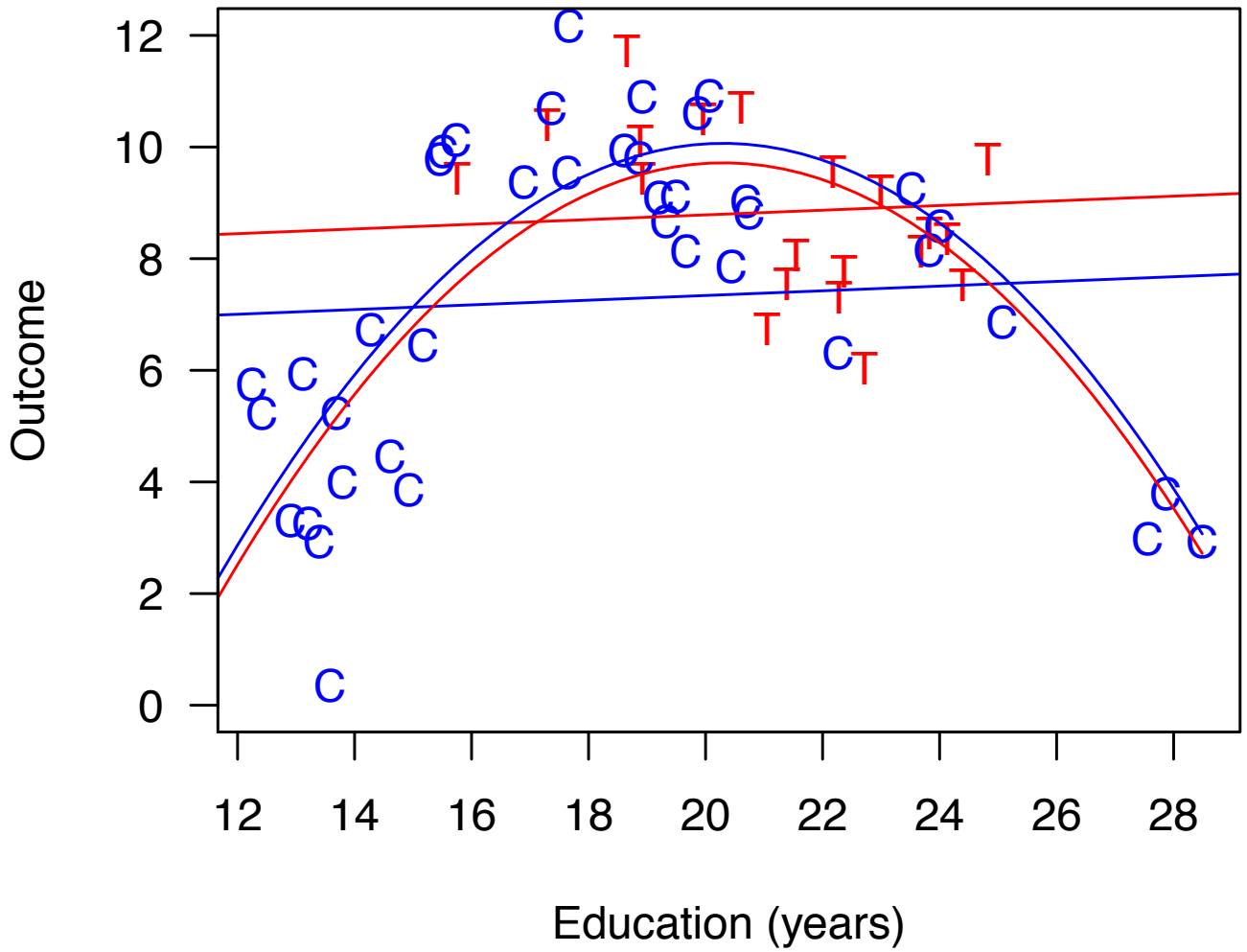
(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)



(slides by King and Nielsen)

Matching to Reduce Model Dependence

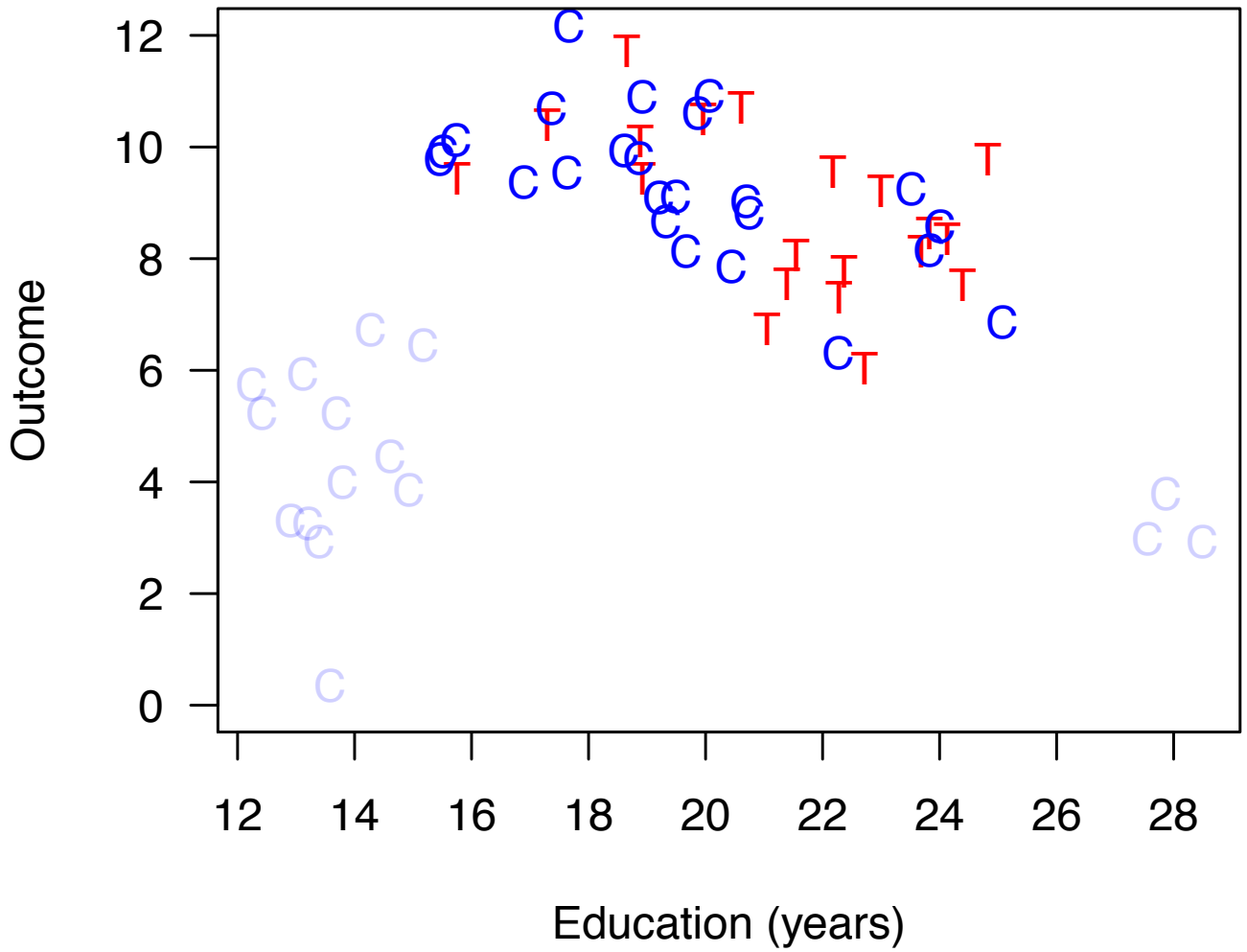
(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)



(slides by King and Nielsen)

Matching to Reduce Model Dependence

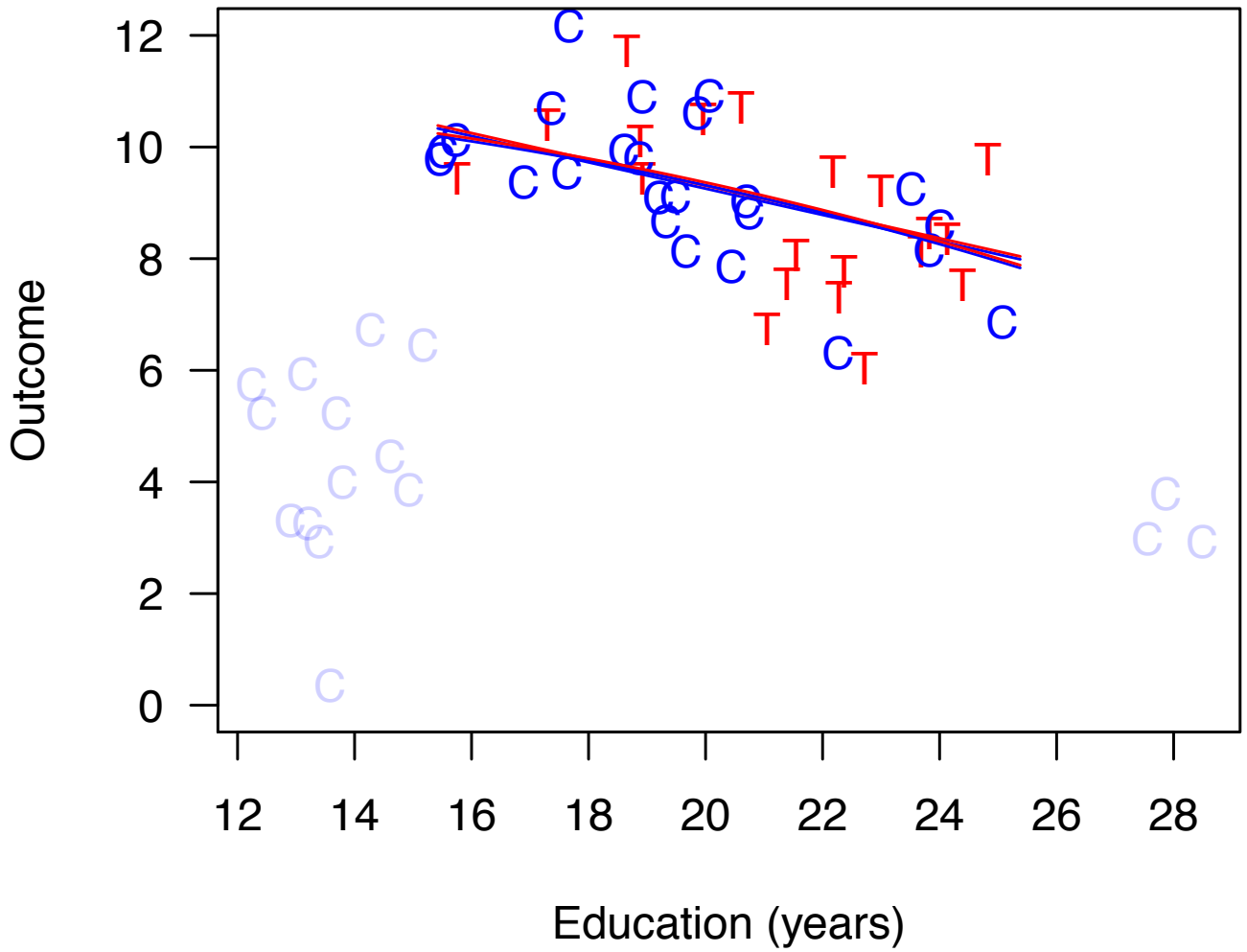
(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)



(slides by King and Nielsen)

Matching to Reduce Model Dependence

(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)



3/23

(slides by King and Nielsen)

King and Nielsen

- Argument 2
  - PSM approximates complete randomization.
  - Better are matching approaches that approximate fully blocked randomization, such as Mahalanobis matching (because complete randomization is less efficient than fully blocked randomization).

Types of Experiments

Balance Covariates:	*Complete Randomization*	*Fully Blocked*
Observed	On average	Exact
Unobserved	On average	On average

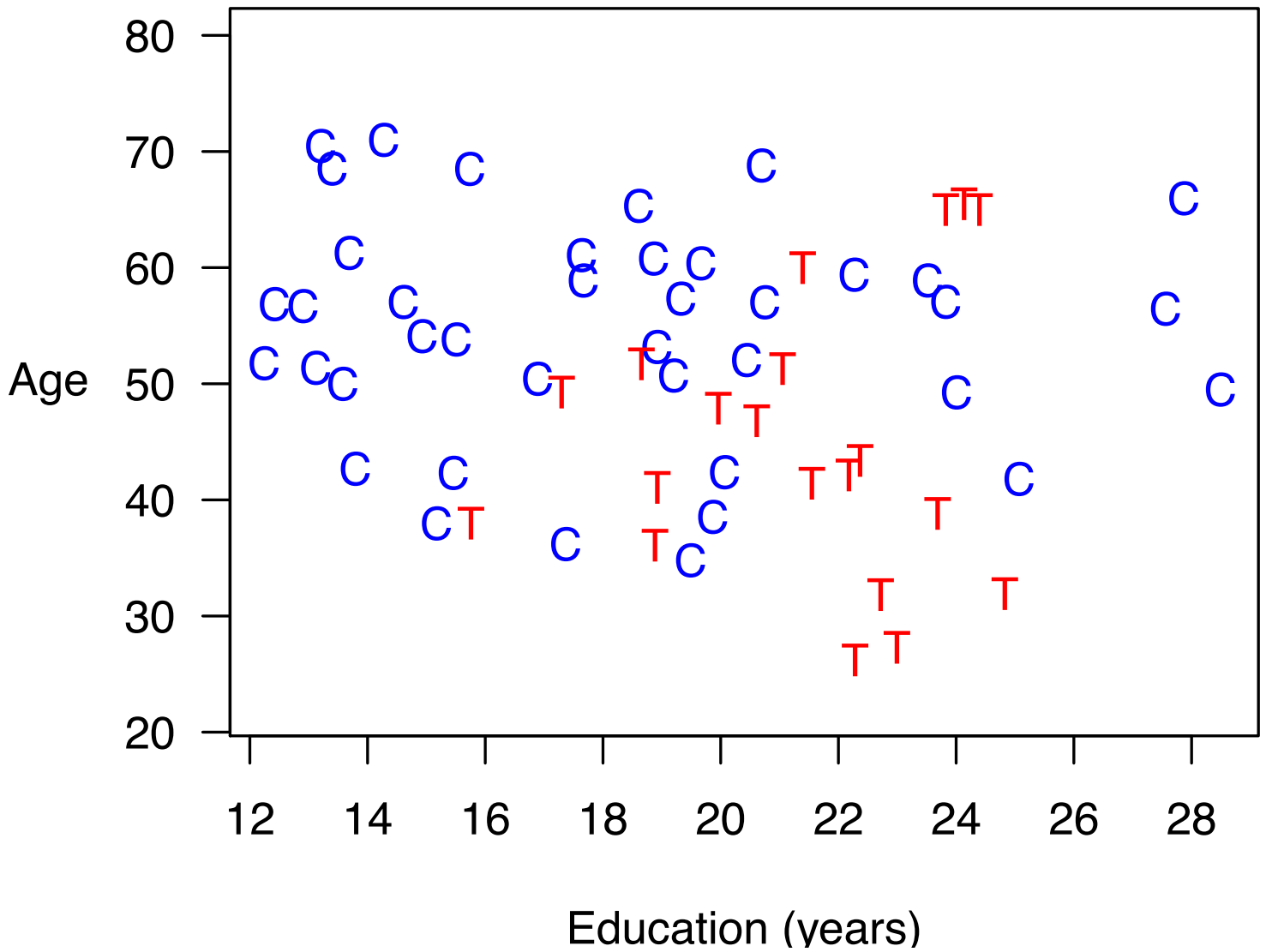
$\rightsquigarrow$ *Fully blocked* dominates *complete randomization* for: imbalance, model dependence, power, efficiency, bias, research costs, robustness. E.g., Imai, King, Nall 2009: SEs 600% smaller!

Goal of Each Matching Method (in Observational Data)

- PSM: *complete randomization*
- Other methods: *fully blocked*
- Other matching methods dominate PSM (wait, it gets worse)

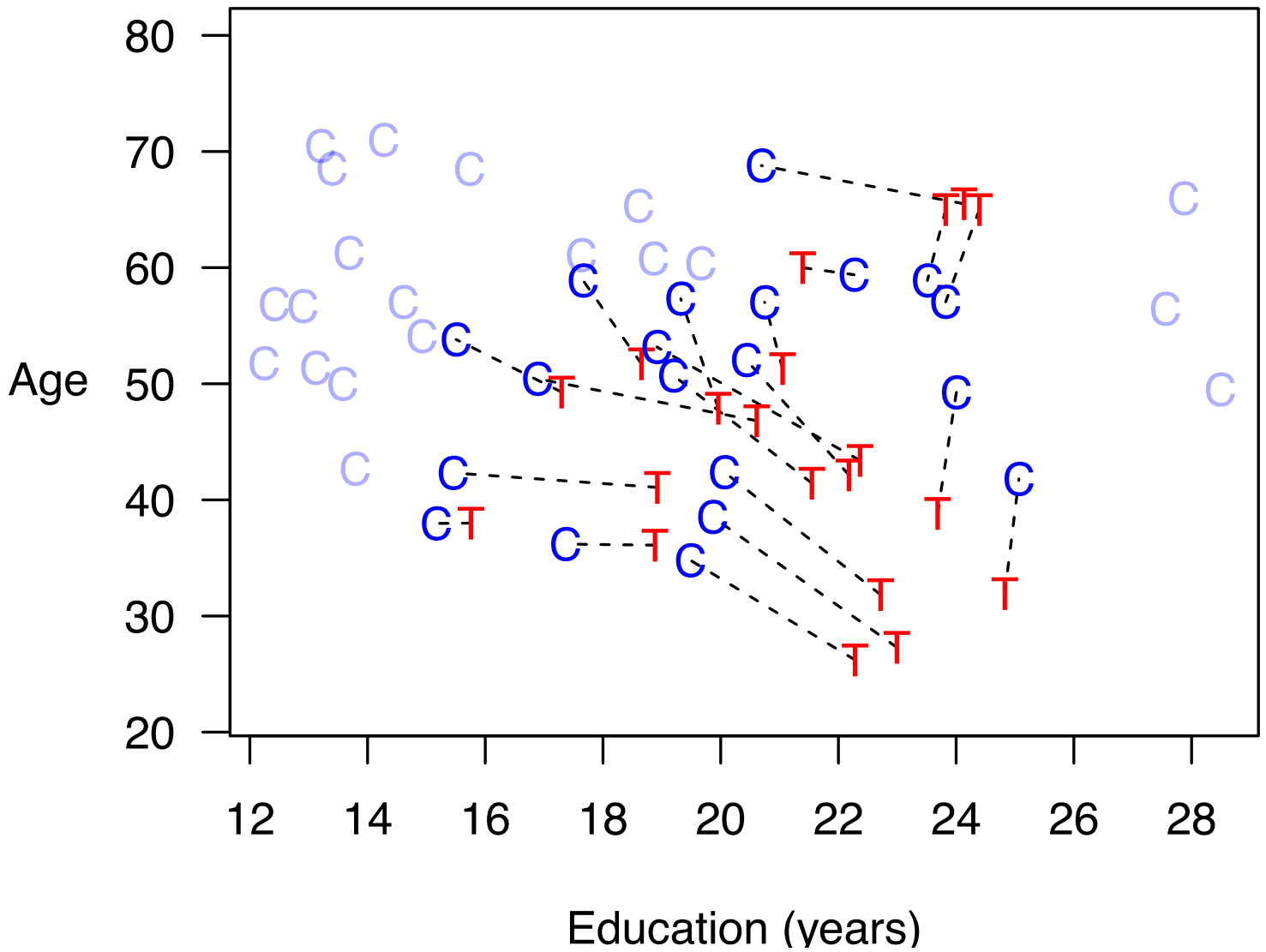
(slides by King and Nielsen)

Mahalanobis Distance Matching



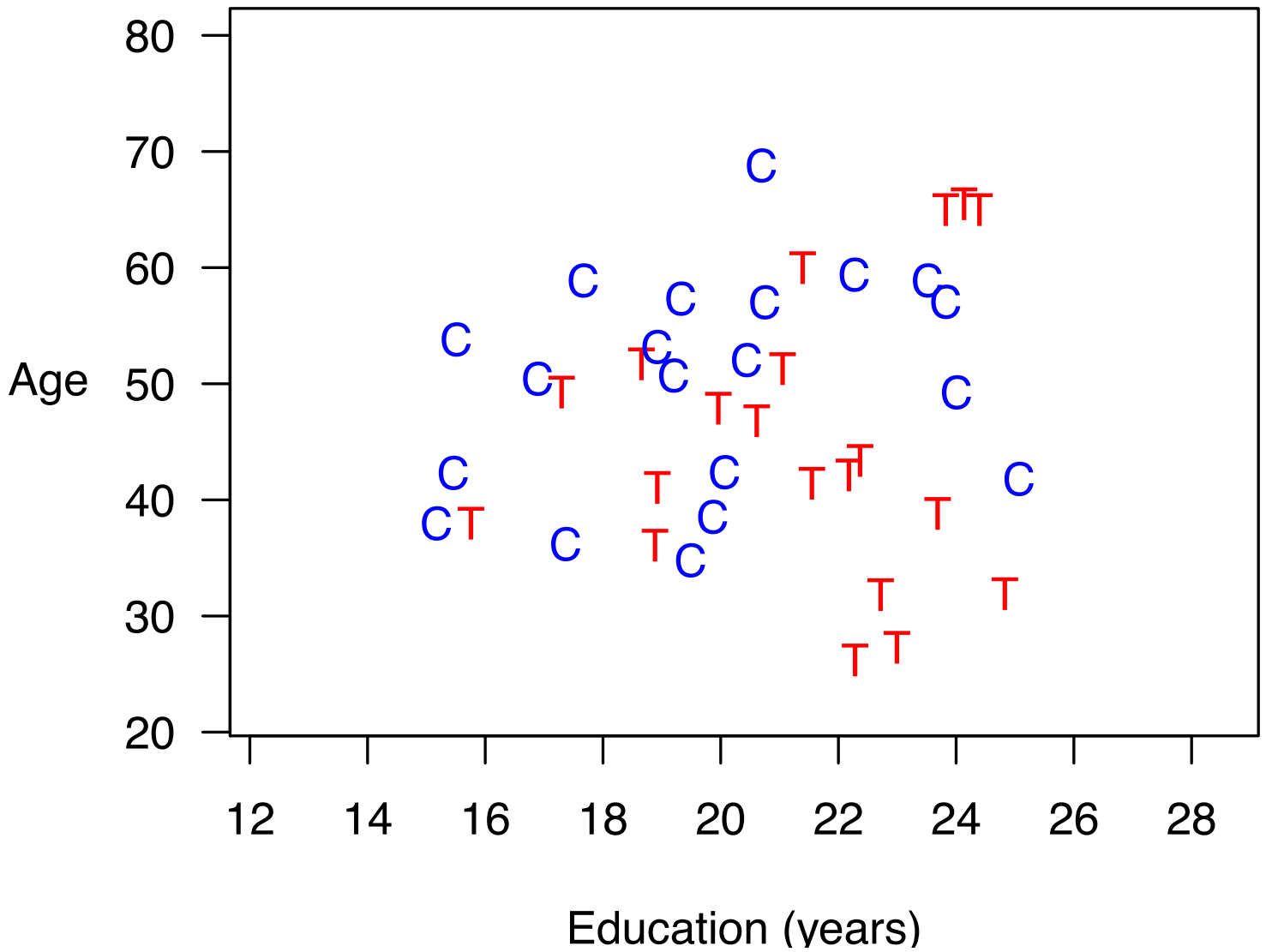
(slides by King and Nielsen)

Mahalanobis Distance Matching



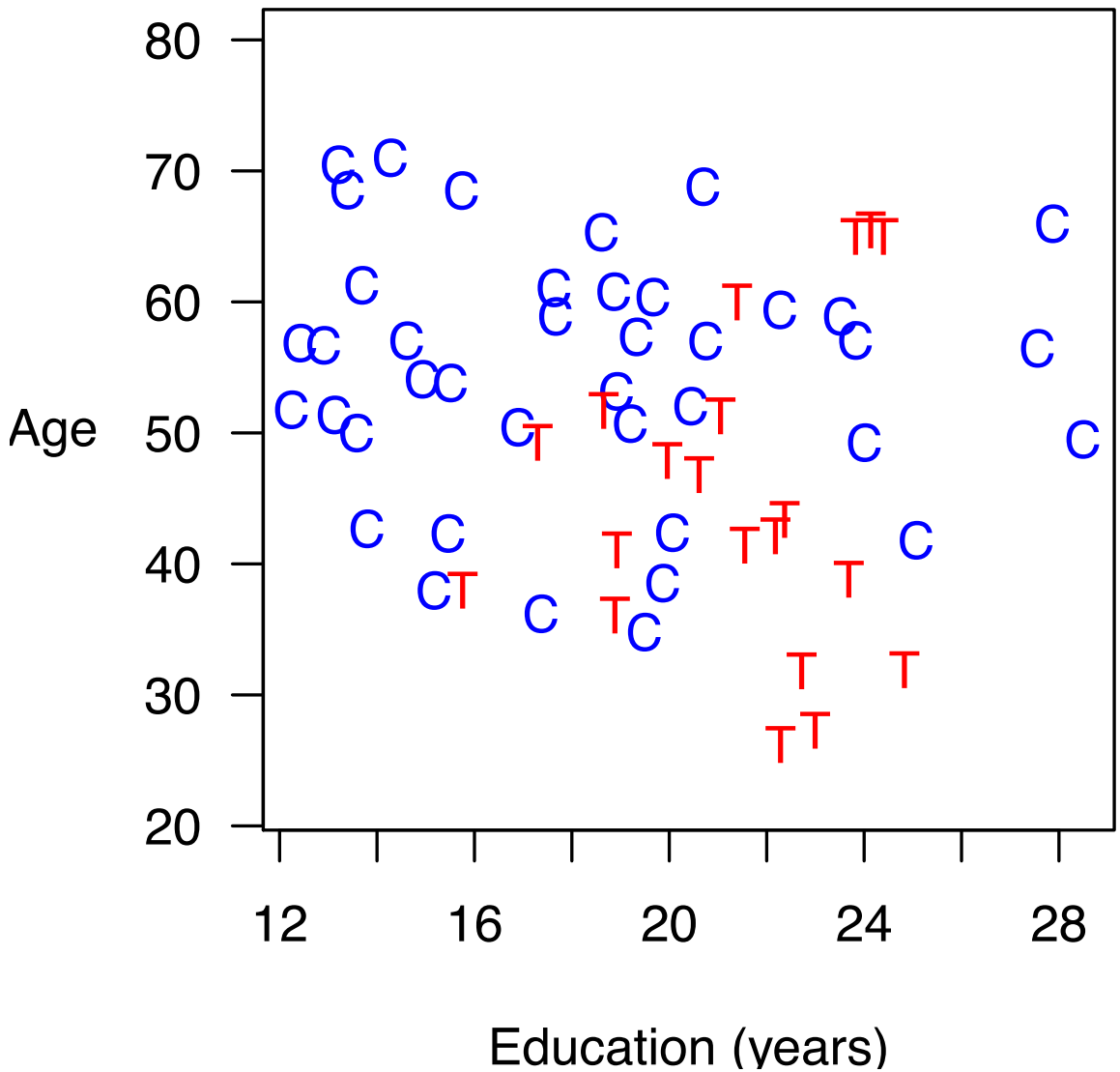
(slides by King and Nielsen)

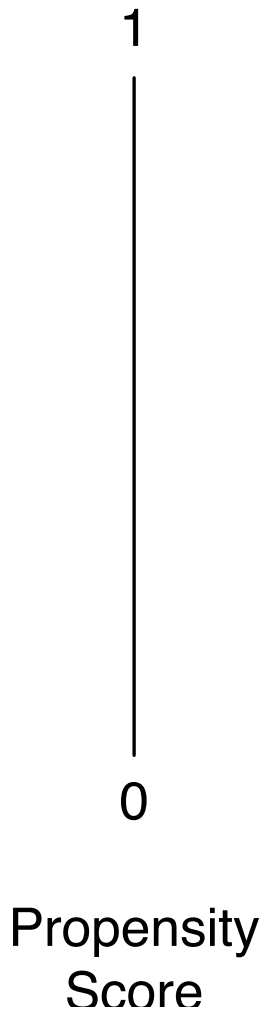
Mahalanobis Distance Matching



(slides by King and Nielsen)

Propensity Score Matching

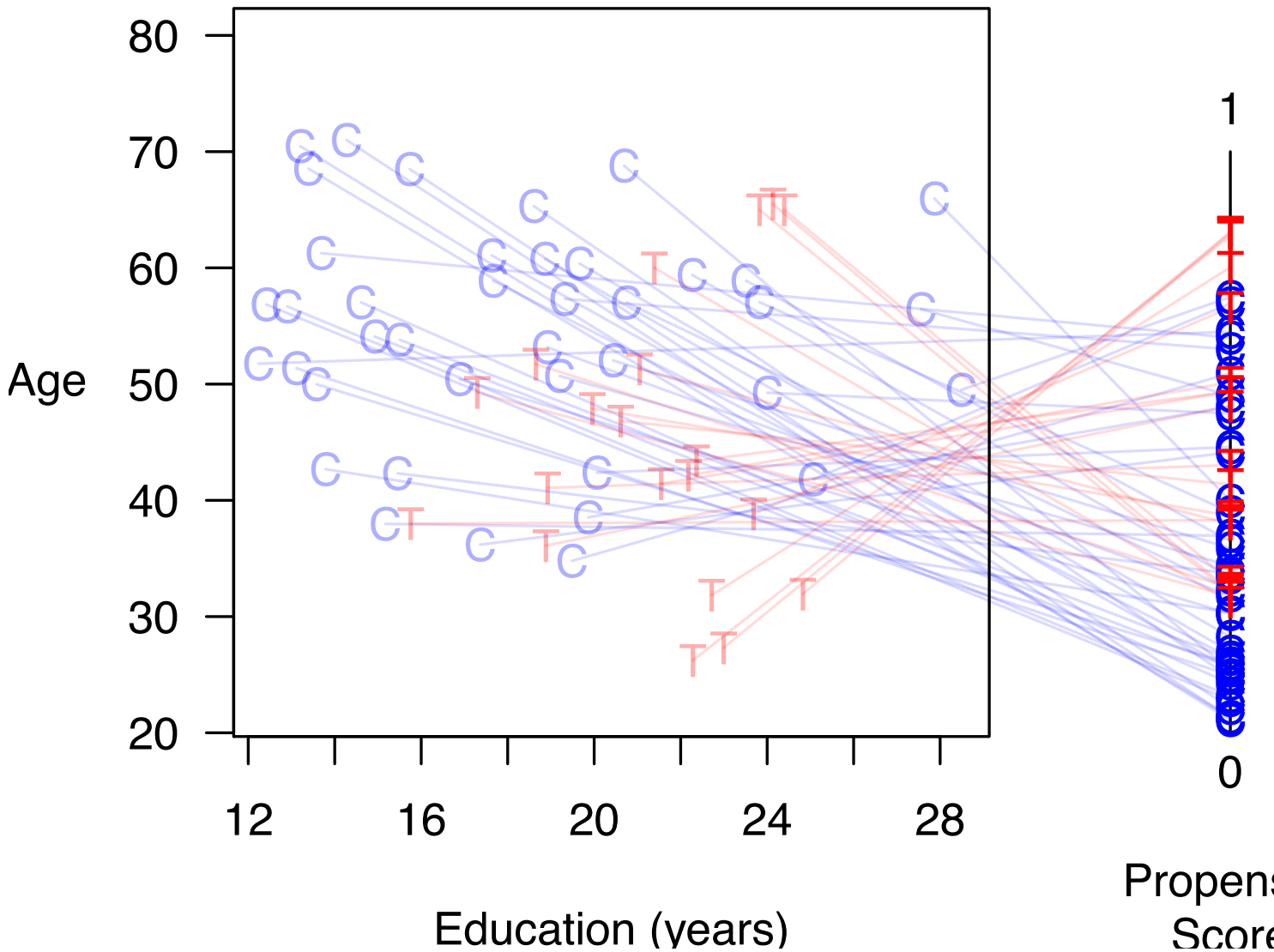




(slides by King and Nielsen)

14/23

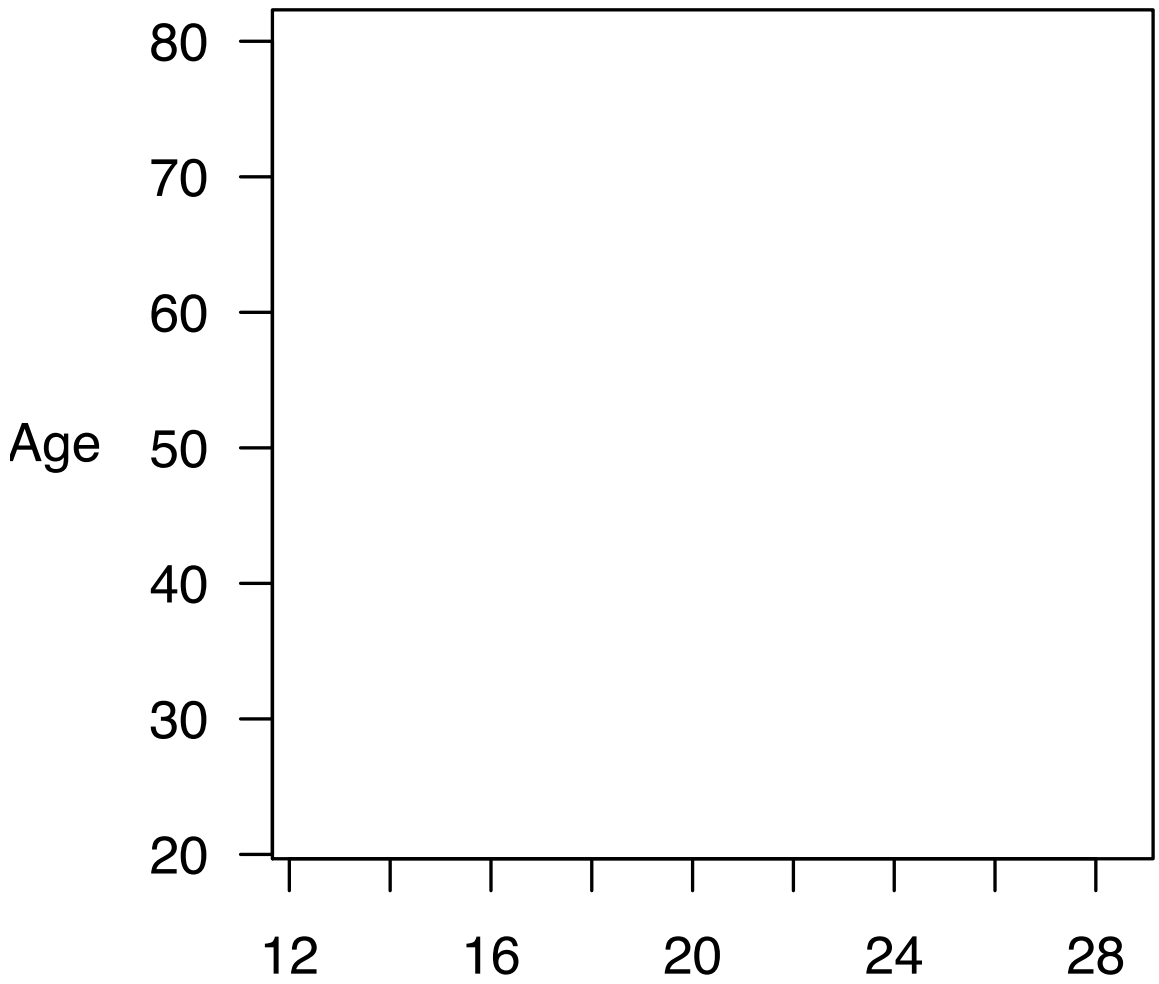
Propensity Score Matching

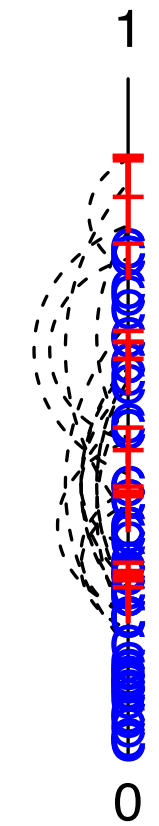


(slides by King and Nielsen)

14/23

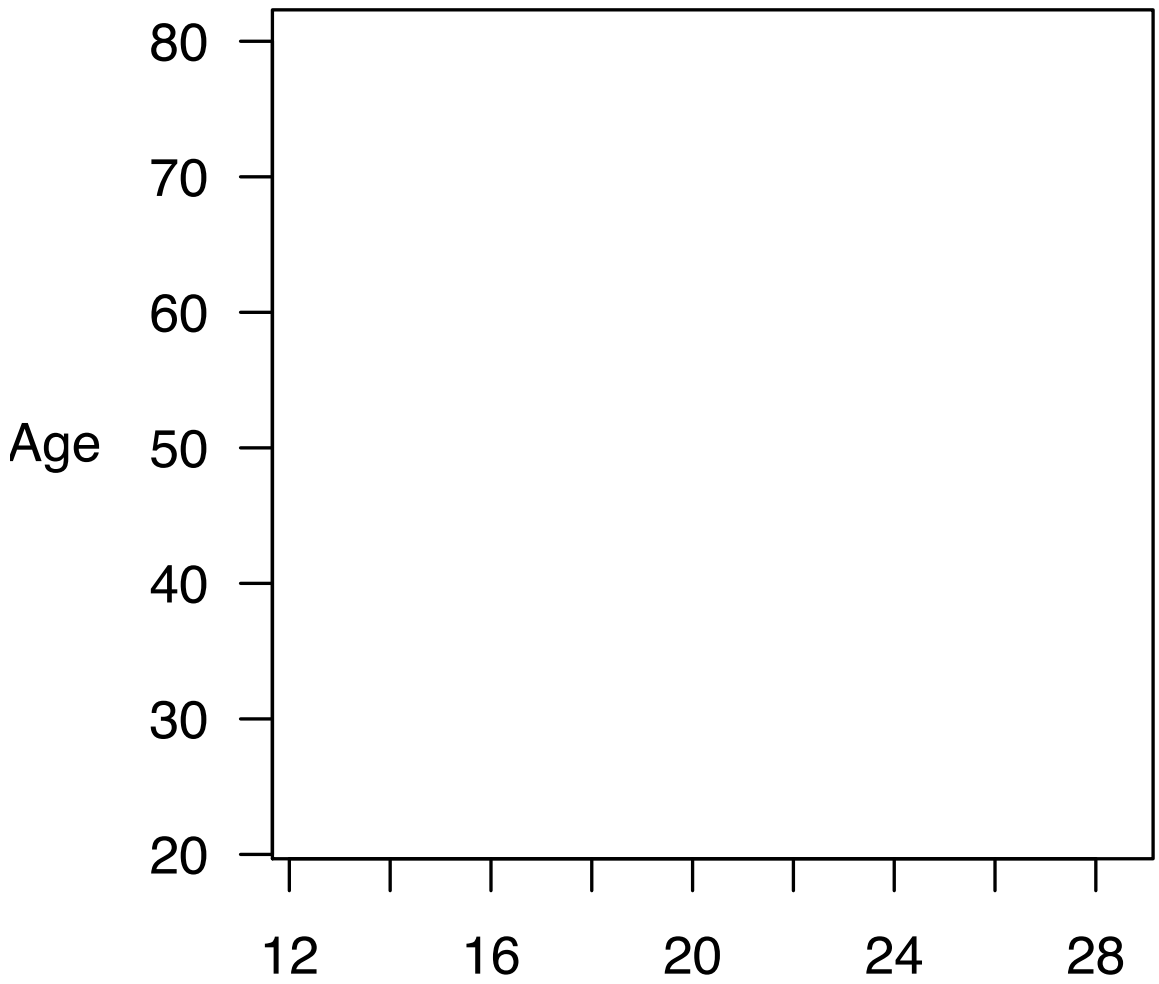
Propensity Score Matching





(slides by King and Nielsen)

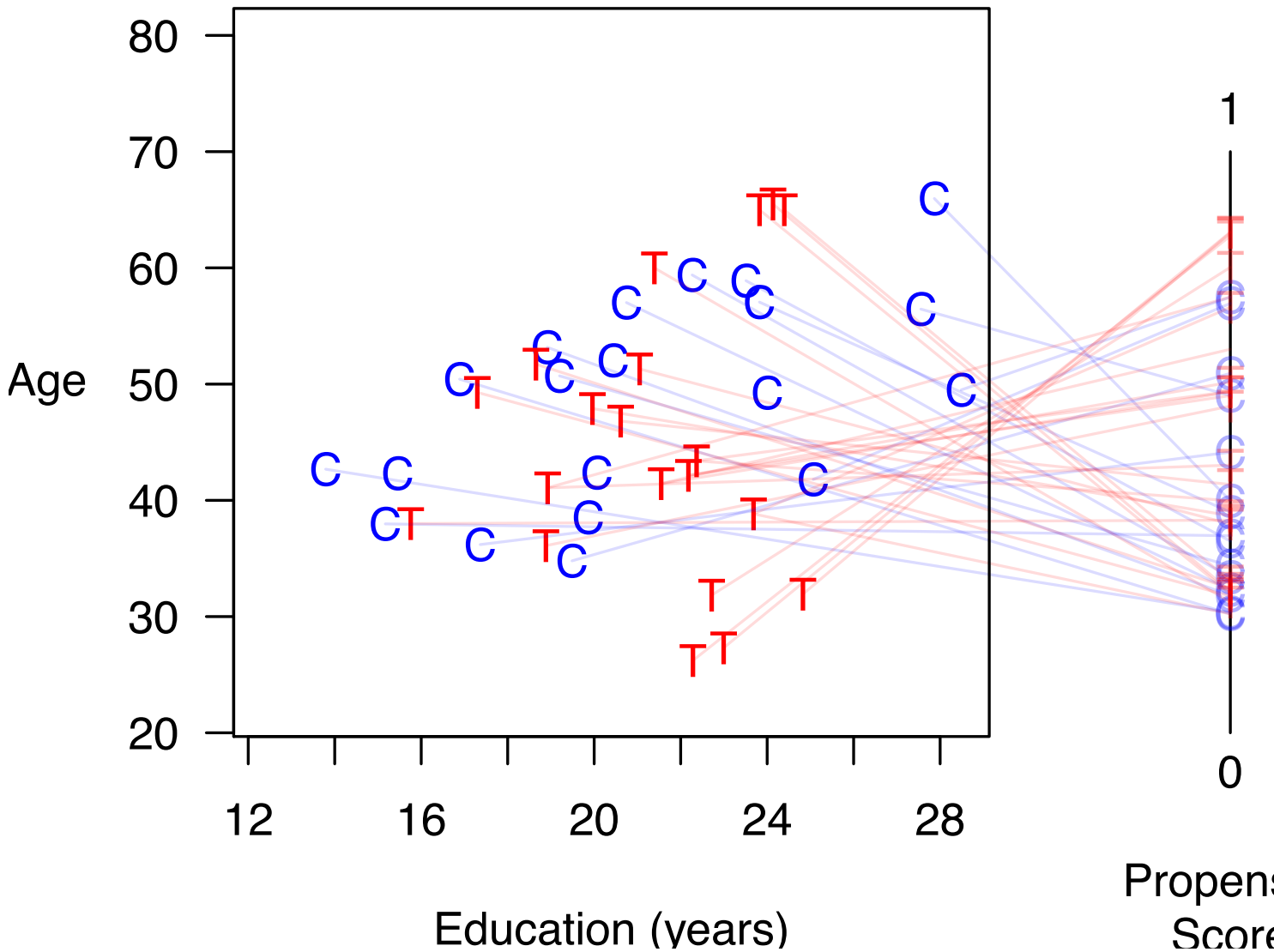
Propensity Score Matching



(slides by King and Nielsen)

14/23

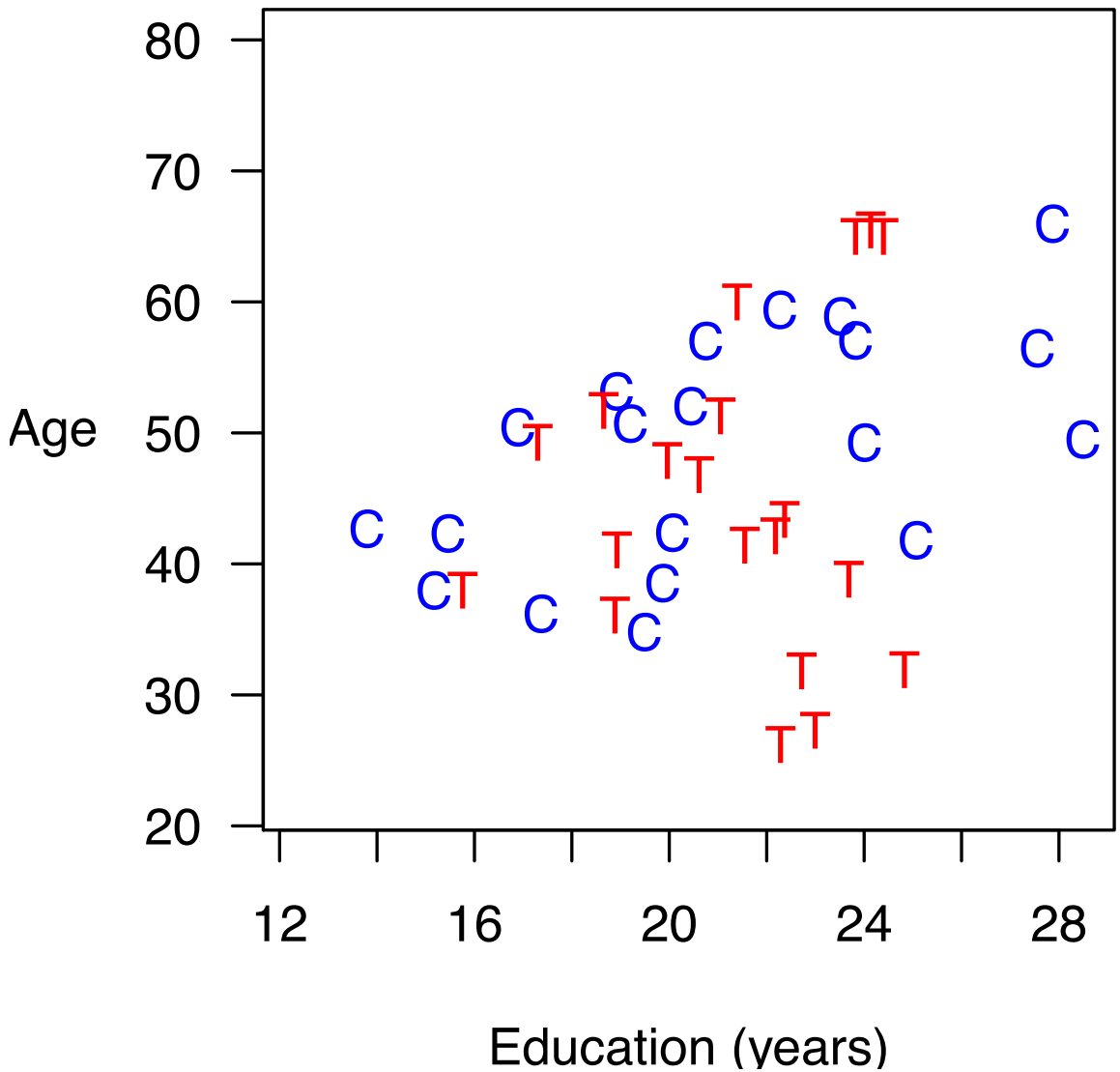
Propensity Score Matching



(slides by King and Nielsen)

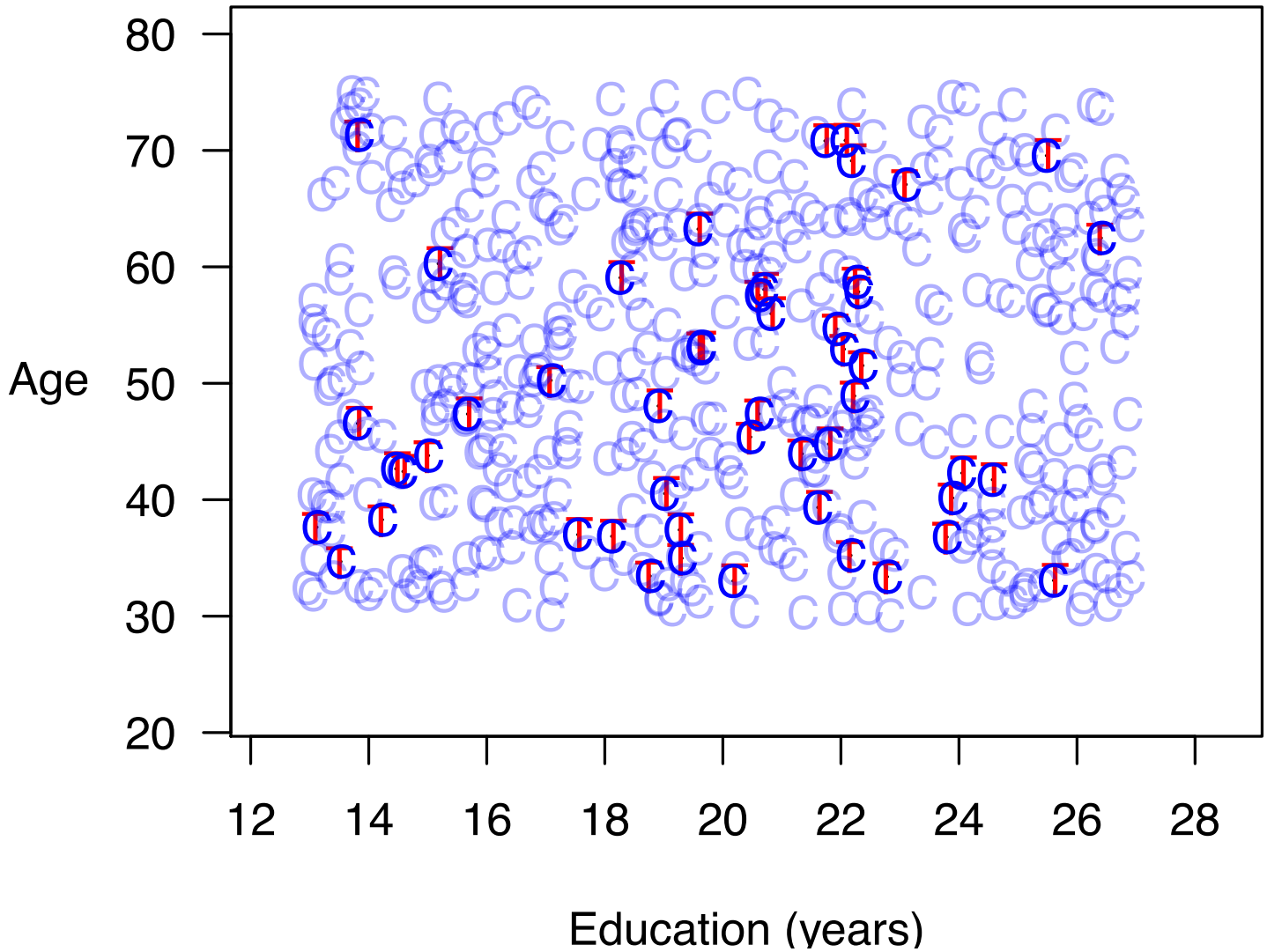
14/23

Propensity Score Matching



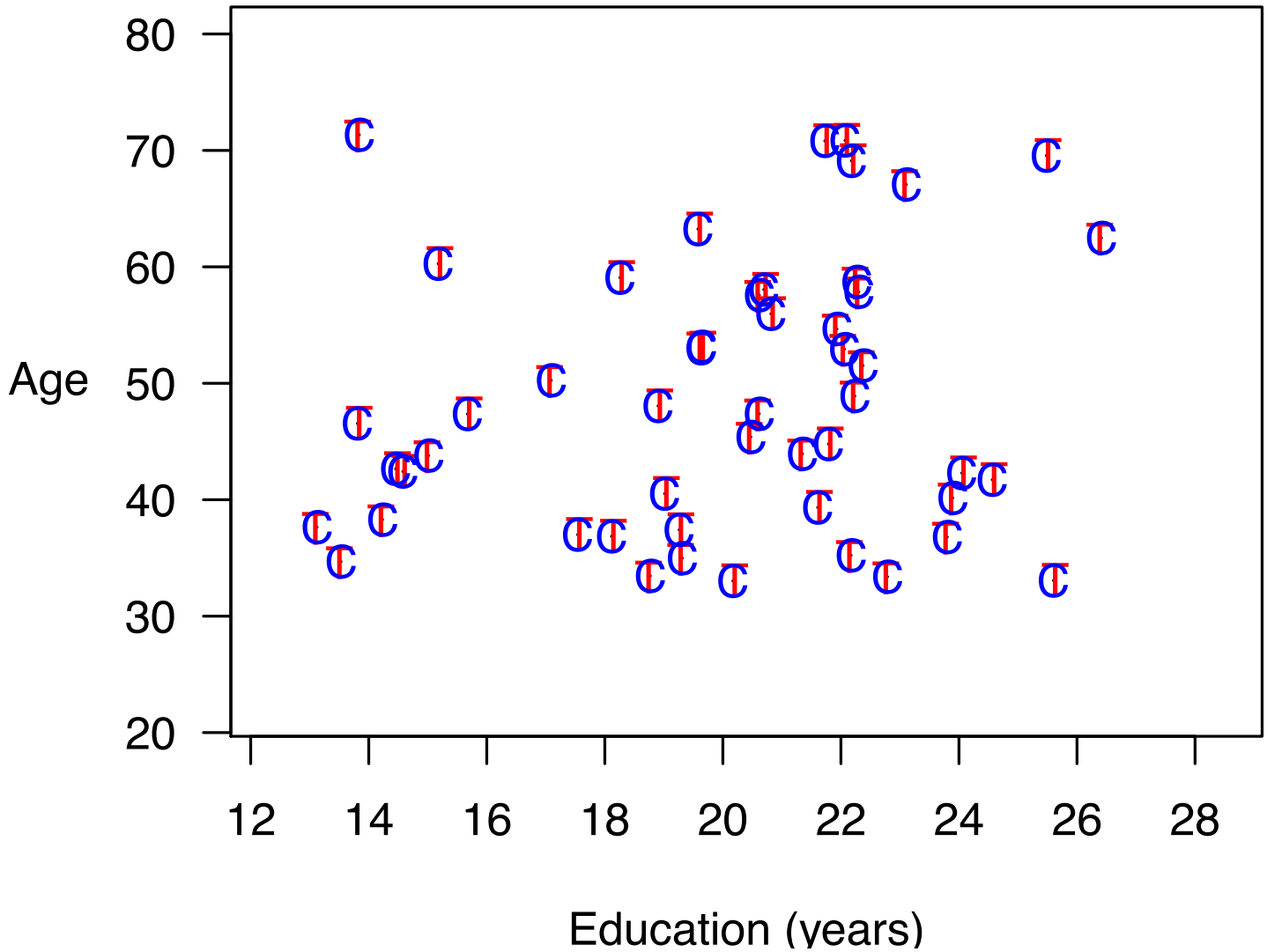
(slides by King and Nielsen)

Best Case: Mahalanobis Distance Matching



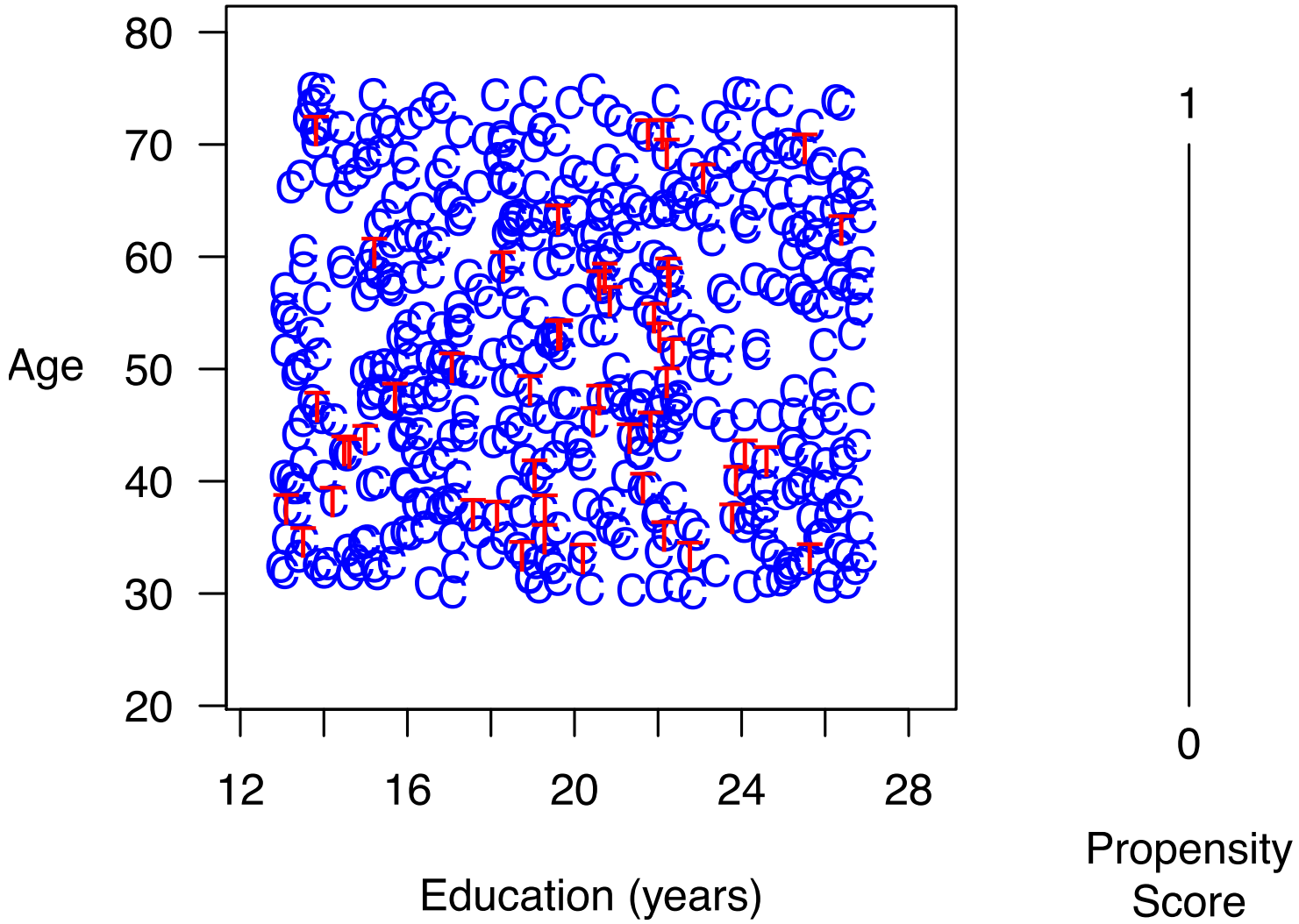
(slides by King and Nielsen)

Best Case: Mahalanobis Distance Matching



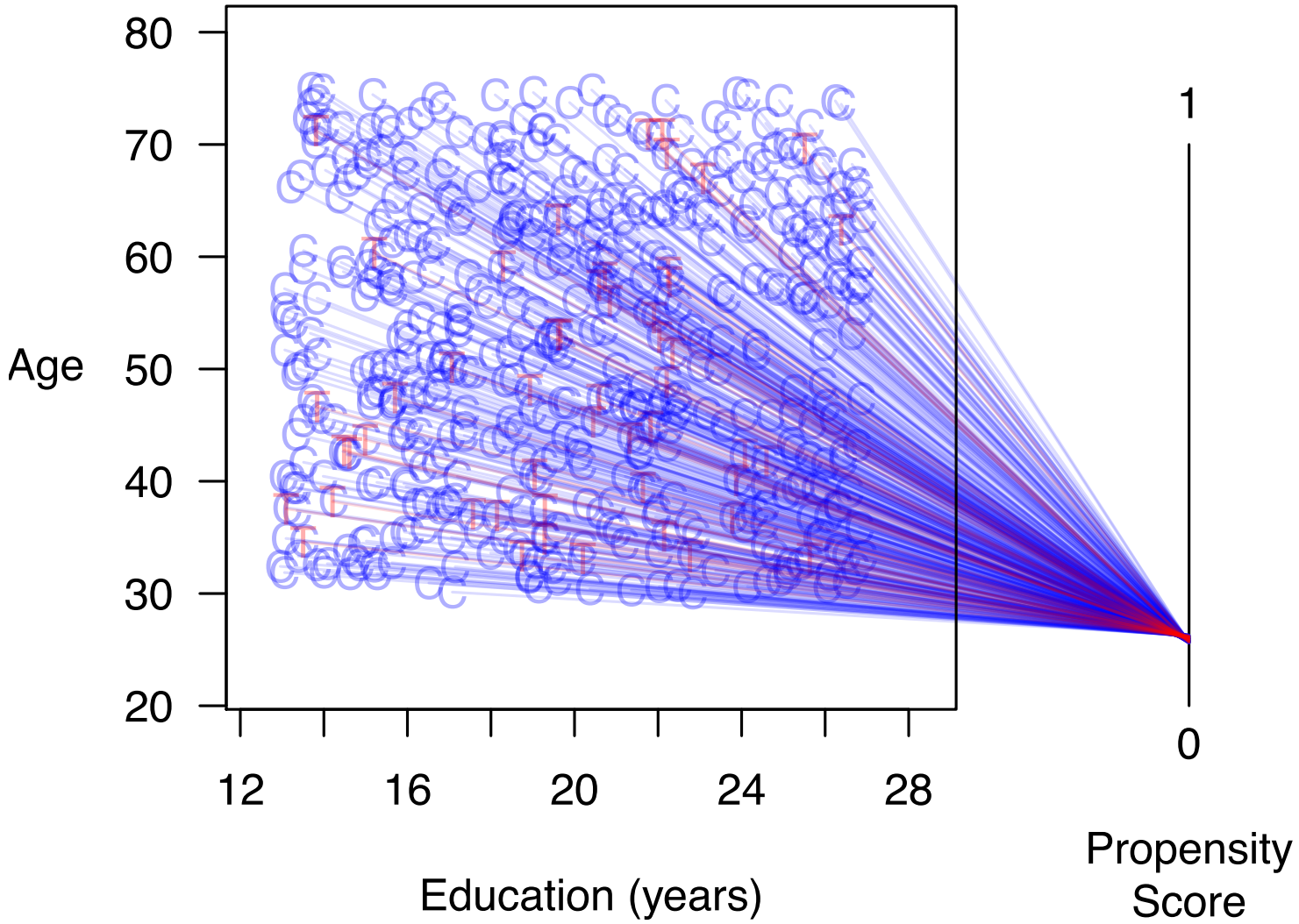
(slides by King and Nielsen)

Best Case: Propensity Score Matching



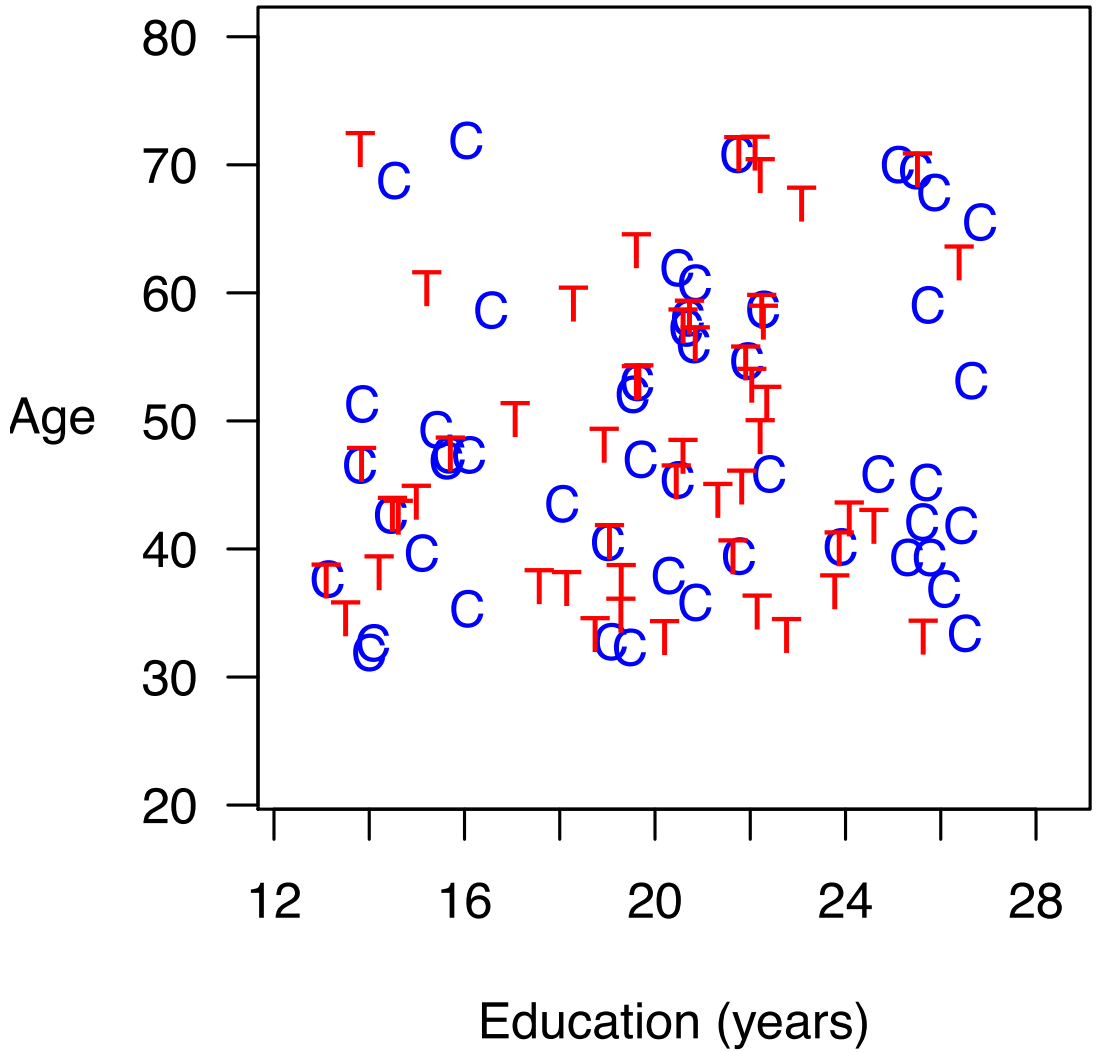
(slides by King and Nielsen)

Best Case: Propensity Score Matching



(slides by King and Nielsen)

Best Case: Propensity Score Matching is Suboptimal



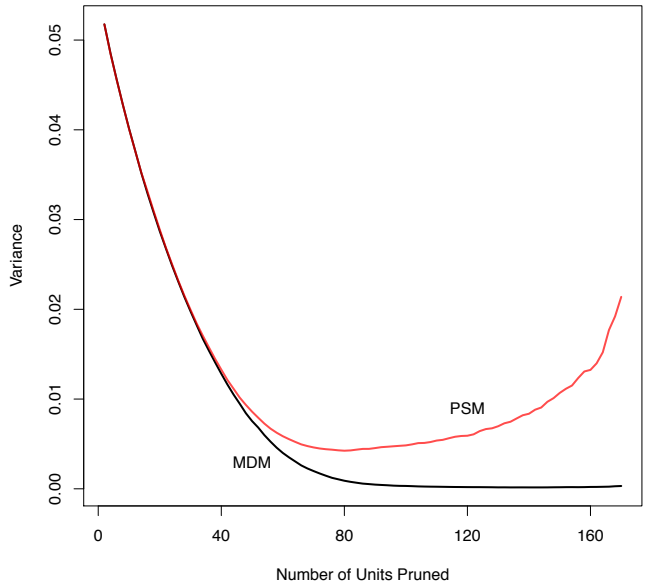
(slides by King and Nielsen)

King and Nielsen

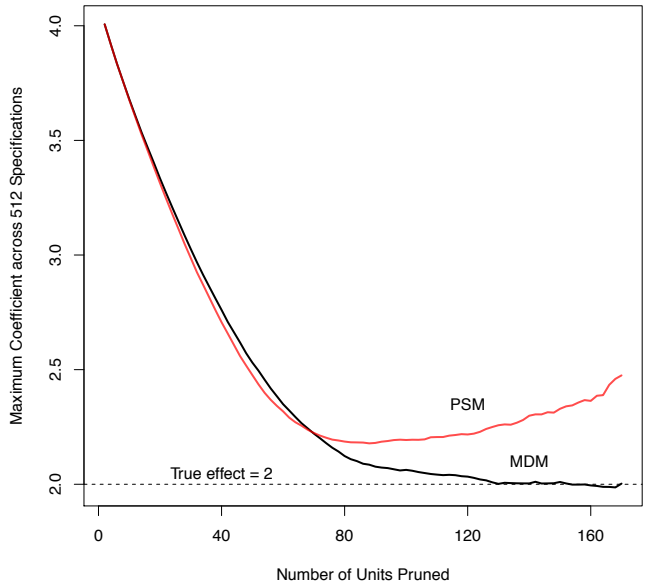
- Argument 3
  - Random pruning (deleting observations at random) increases imbalance. This is because the sample size decreases so that variance increases (large differences become more likely).
  - More imbalance/variance means more model dependence and researcher discretion.
  - Because PSM approximates complete randomization, it engages in random pruning.
  - PSM Paradox ("when you do 'better,' you do worse")
    - ★ When matching is made more strict (e.g., by decreasing the size of the caliper) PSM, like other matching methods, typically reduces imbalance. But soon the PSM Paradox kicks in, such that further pruning quickly increases imbalance.
    - ★ If the data is such that there are no big differences between treated and untreated to begin with, the PSM Paradox kicks in almost immediately.

PSM Increases Model Dependence & Bias

Model Dependence



Bias

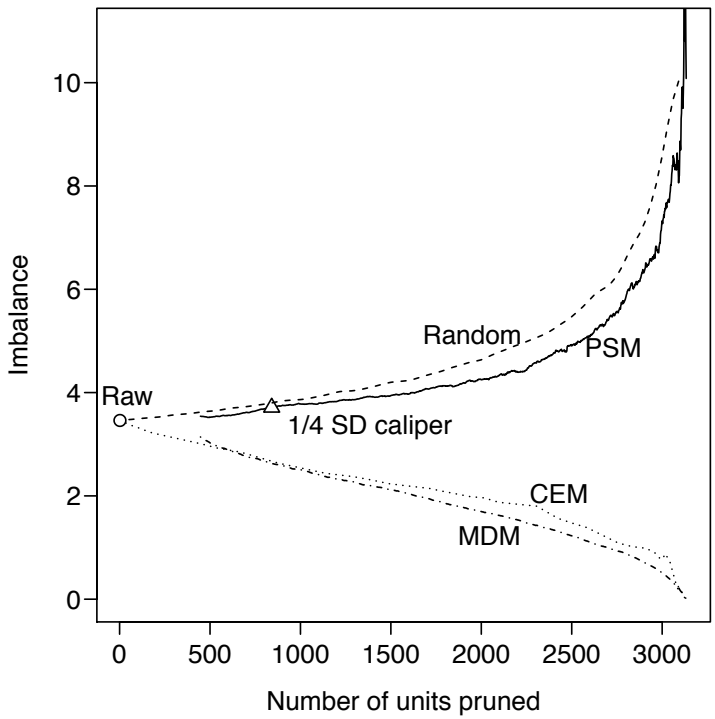


$$Y_i = 2T_i + X_{1i} + X_{2i} + \epsilon_i$$
$$\epsilon_i \sim N(0, 1)$$

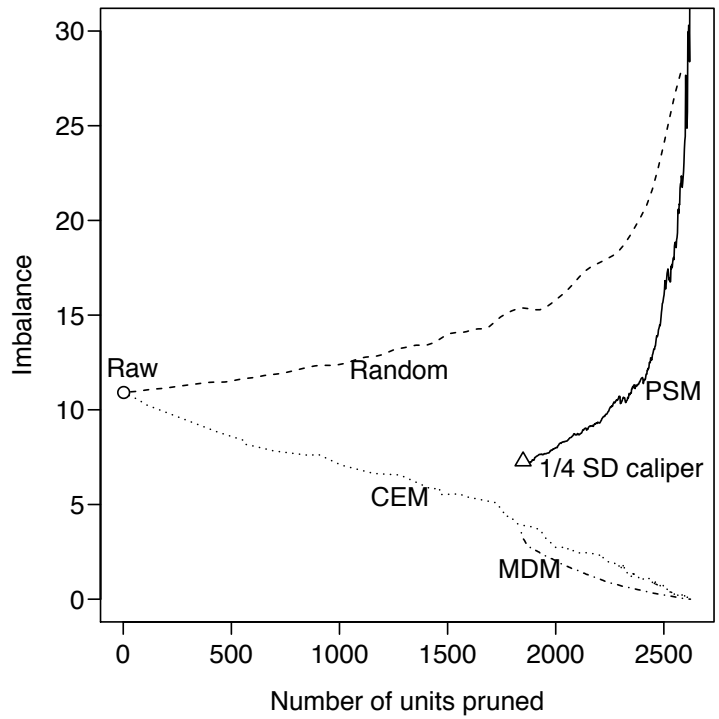
(slides by King and Nielsen)

The Propensity Score Paradox in Real Data

Finkel et al. (JOP, 2012)



Nielsen et al. (AJPS, 2011)



Similar pattern for $>$ 20 other real data sets we checked

(slides by King and Nielsen)

1 Potential Outcomes and Causal Inference

2 Matching

3 Propensity Score Matching

4 King and Nielsen's "Why Propensity Scores Should Not Be Used for Matching"

5 Are King and Nielsen right?

6 Illustration using `kmatch`

7 Conclusions

Are King and Nielsen right?

- Argument 1
  - Model dependence (i.e. dependence of results on modeling decisions made by the researcher) is bad because it leads to bias (people are selective in their choices even if they try not to be).
  - Matching is good because it reduces model dependence.
- I fully agree!
- My view, however, may be somewhat less pessimistic. I believe that research results can be credible if researchers are well educated so that they know what they are doing and if modeling decisions are made transparent and robustness of results is evaluated (and documented).

Are King and Nielsen right?

- Argument 2
  - PSM approximates complete randomization.
  - Better are matching approaches that approximate fully blocked randomization, such as Mahalanobis matching (because complete randomization is less efficient than fully blocked randomization).
- That complete randomization is less efficient than fully blocked randomization – given the sample size – is of course true (how large the efficiency gains are further depends on the strength of the relation between X and Y).
- However, if blocking reduces the sample size, it is not a priori clear whether estimates from the blocked sample are more efficient than estimates from the full sample (although often they will be).

Are King and Nielsen right?

- Argument 2
  - PSM approximates complete randomization.
  - Better are matching approaches that approximate fully blocked randomization, such as Mahalanobis matching (because complete randomization is less efficient than fully blocked randomization).
- That PSM approximates complete randomization is only partially true. PSM approximates complete randomization *within observations with the same propensity score*. Hence, PSM is somewhere between complete randomization and fully blocked randomization.
  - If the X variables have no relation to T (treatment), then all observations have the same propensity score. Hence we end up with complete randomization.
  - If the X variables have a strong effect on T, there is lots of blocking.

Are King and Nielsen right?

- Argument 3
  - Random pruning $\Rightarrow$ imbalance $\Rightarrow$ more model dependence.
  - PSM $\Rightarrow$ complete randomization $\Rightarrow$ lots of random pruning.
  - PSM Paradox: "when you do 'better,' you do worse"
- That random pruning makes things worse is, of course, true because it unnecessarily reduces the sample size (without changing anything else).
- As argued above, that PSM applies random pruning is only true for X variables unrelated to T (so that we are in a "local" complete randomization situation; although something similar can probably also happen if effects from several X's cancel each other out).
- Furthermore, it is only true if you employ a matching algorithm that throws away good matches! King and Nielsen's results seem to be based on the worst possible algorithm: one-to-one matching without replacement.

Are King and Nielsen right?

- Argument 3
  - Random pruning ⇒ imbalance ⇒ more model dependence.
  - PSM ⇒ complete randomization ⇒ lots of random pruning.
  - PSM Paradox: "when you do 'better,' you do worse"
- If you use a matching algorithm that does not throw away good matches, such as radius or kernel matching (or also nearest-neighbor matching as long as all ties are kept and observations are matched with replacement), no random pruning is applied.
  - Such algorithms block (and hence prune) where it is necessary to prevent bias, but they average where such pruning is not necessary.
  - Hence, efficiency differences between PSM and multivariate matching should only be minor for such algorithms.

Are King and Nielsen right?

- Argument 3
  - Random pruning $\Rightarrow$ imbalance $\Rightarrow$ more model dependence.
  - PSM $\Rightarrow$ complete randomization $\Rightarrow$ lots of random pruning.
  - PSM Paradox: "when you do 'better,' you do worse"
- True is that post-matching modeling can do more harm with PSM than with multivariate matching (because PSM leaves more "free" variance in X that can exploited by modeling decisions).
- In general, post-matching analyses are more limited for PSM than for multivariate matching. For example, results from subgroup analyses may not be valid (you'd need to apply PSM stratified by subgroups in this case).

1 Potential Outcomes and Causal Inference

2 Matching

3 Propensity Score Matching

4 King and Nielsen's "Why Propensity Scores Should Not Be Used for Matching"

5 Are King and Nielsen right?

6 Illustration using `kmatch`

7 Conclusions

Illustration using kmatch

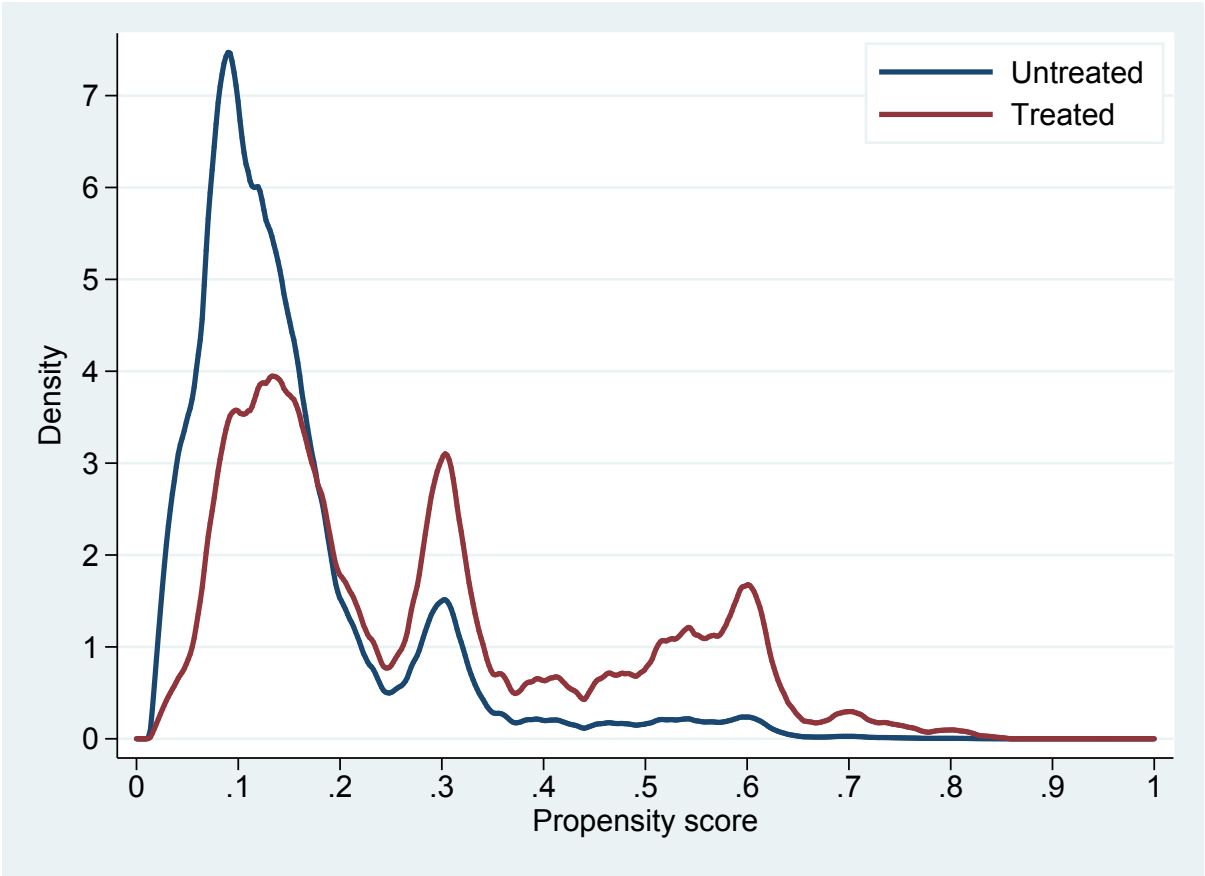
- `kmatch`: new matching software for Stata that has been written over the last few months; available from SSC (`ssc install kmatch`).
- Some key features:
  - Multivariate Distance Matching (MDM) and Propensity Score Matching (PSM) (or MDM and PSM combined).
  - Optional exact matching.
  - Optional regression-adjustment bias-correction.
  - Kernel matching, ridge matching, or nearest-neighbor matching.
  - Automatic bandwidth selectors for kernel/ridge matching.
  - Flexible specification of scaling matrix for MDM.
  - Joint analysis of multiple subgroups and multiple outcome variables.
  - Various post-estimation commands for balancing and common-support diagnostics.
  - Computationally efficient implementation.

Illustration using kmatch

- Simulation:
  - Population data from Swiss census of 2000.
  - Outcome: Treiman occupational prestige (recoded from ISCO codes of the current job using command `iskotrei` by Hendrickx 2002) (values from 6 to 78; mean 44).
  - Estimand: ATT of nationality on occupational prestige, with resident aliens as the treatment group and Swiss nationals as the control group.
  - Control variables: gender, age, and highest educational degree.
  - Population restricted to people between 24 to 60 years old who are working.
  - 2'308'006 individuals, of which 17.5% belong to the treatment group.
  - Draw random samples (N = 500, 1000, or 5000) from population and compute various matching estimators.

Illustration using kmatch

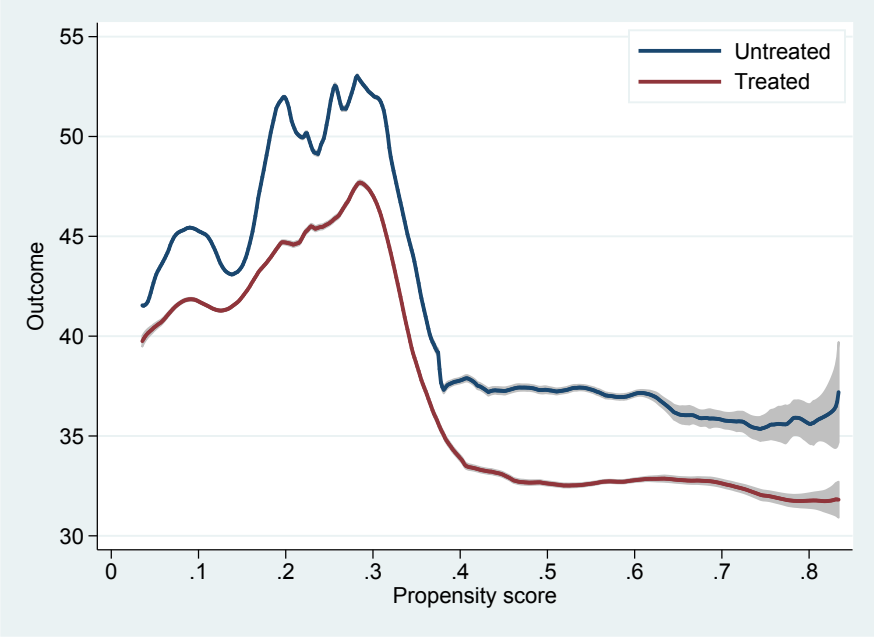
- Substantial differences between resident aliens and Swiss nationals on all three covariates.
- Propensity score in population

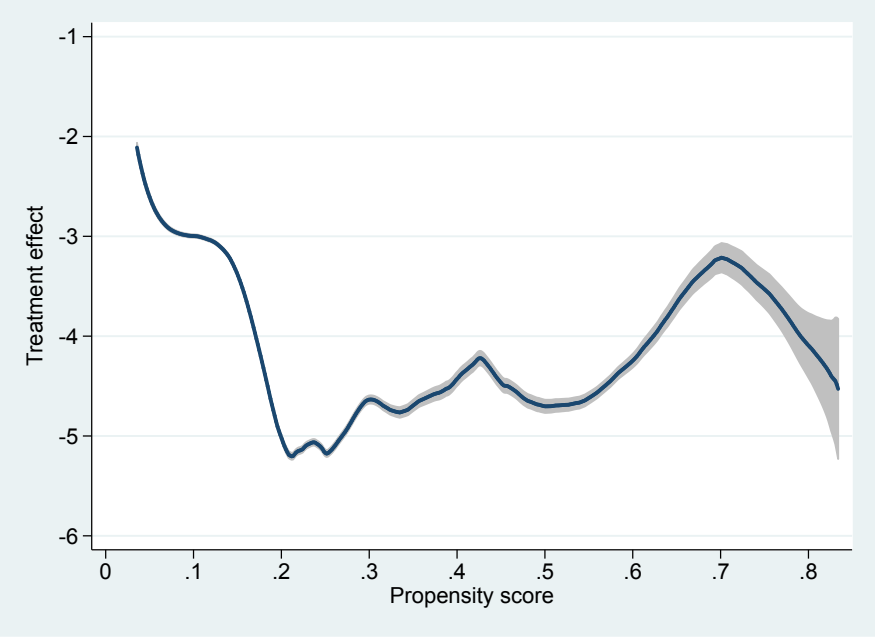


- McFadden $R^2 = 0.121$

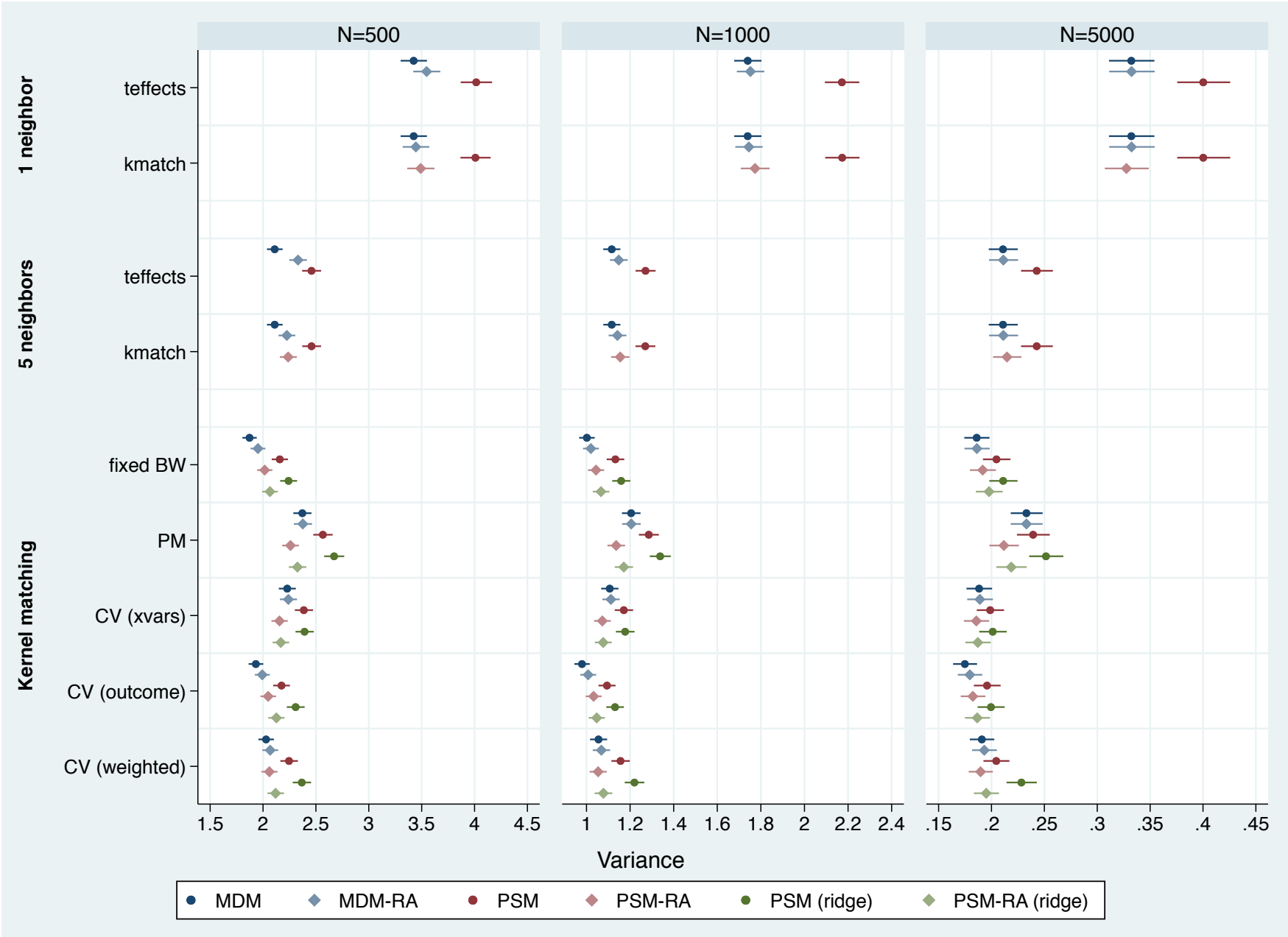
Illustration using kmatch

- Raw mean difference in occupational prestige (NATE): −4.79
- Population ATT (computed from fully stratified data): −3.96
- Some treatment effect heterogeneity (ATE = −3.51, ATC = −3.41)

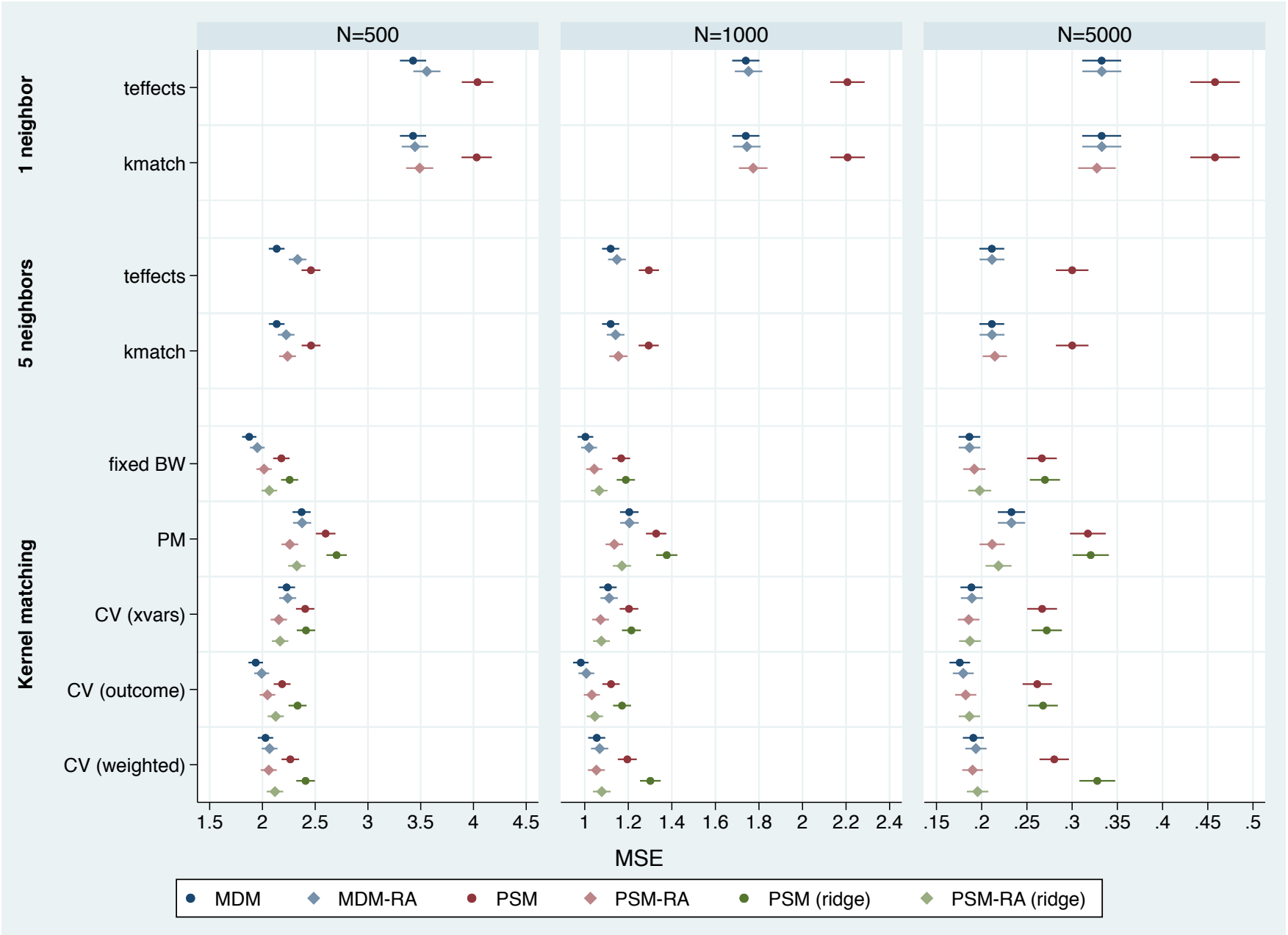




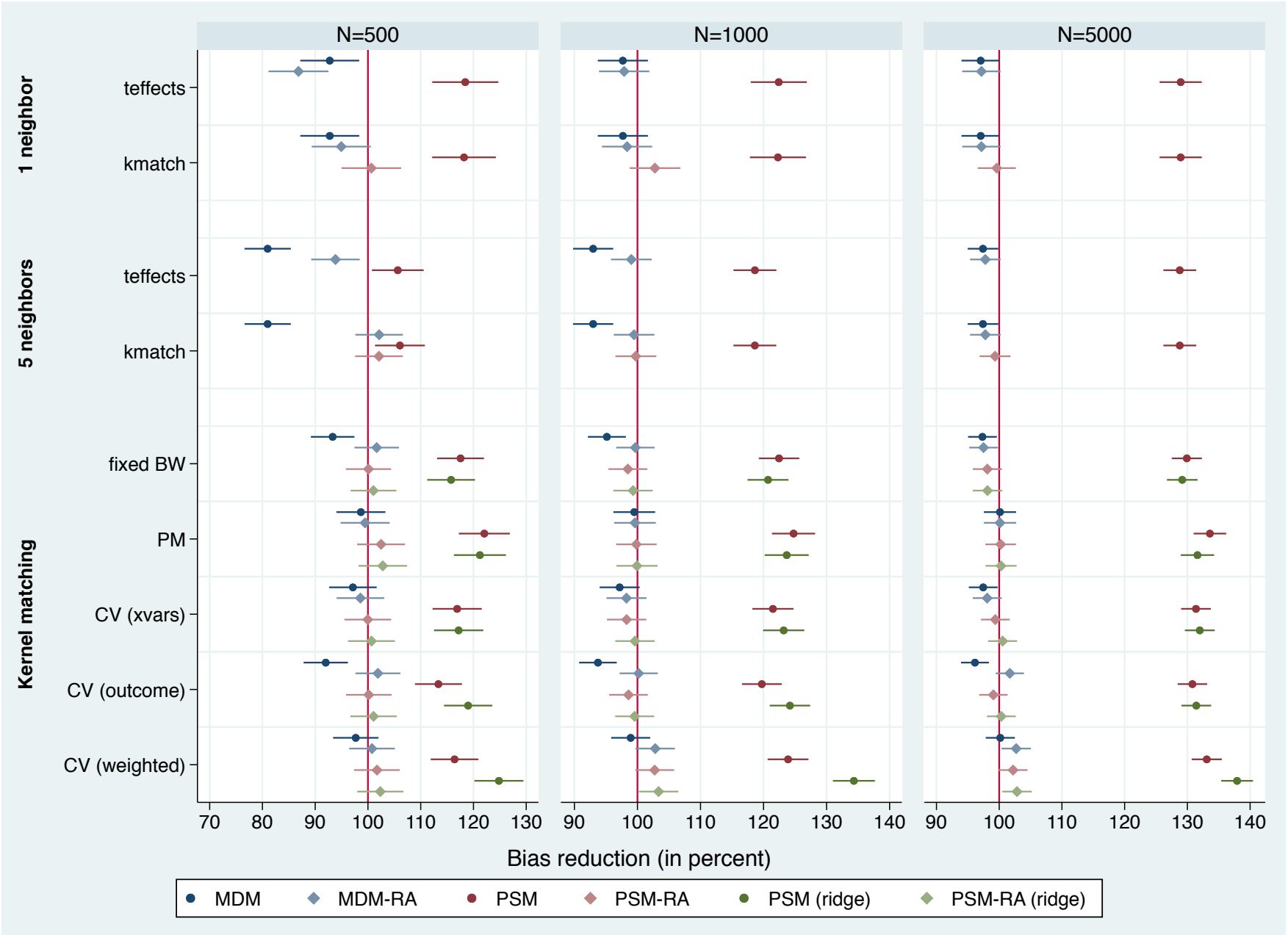
Results: Variance



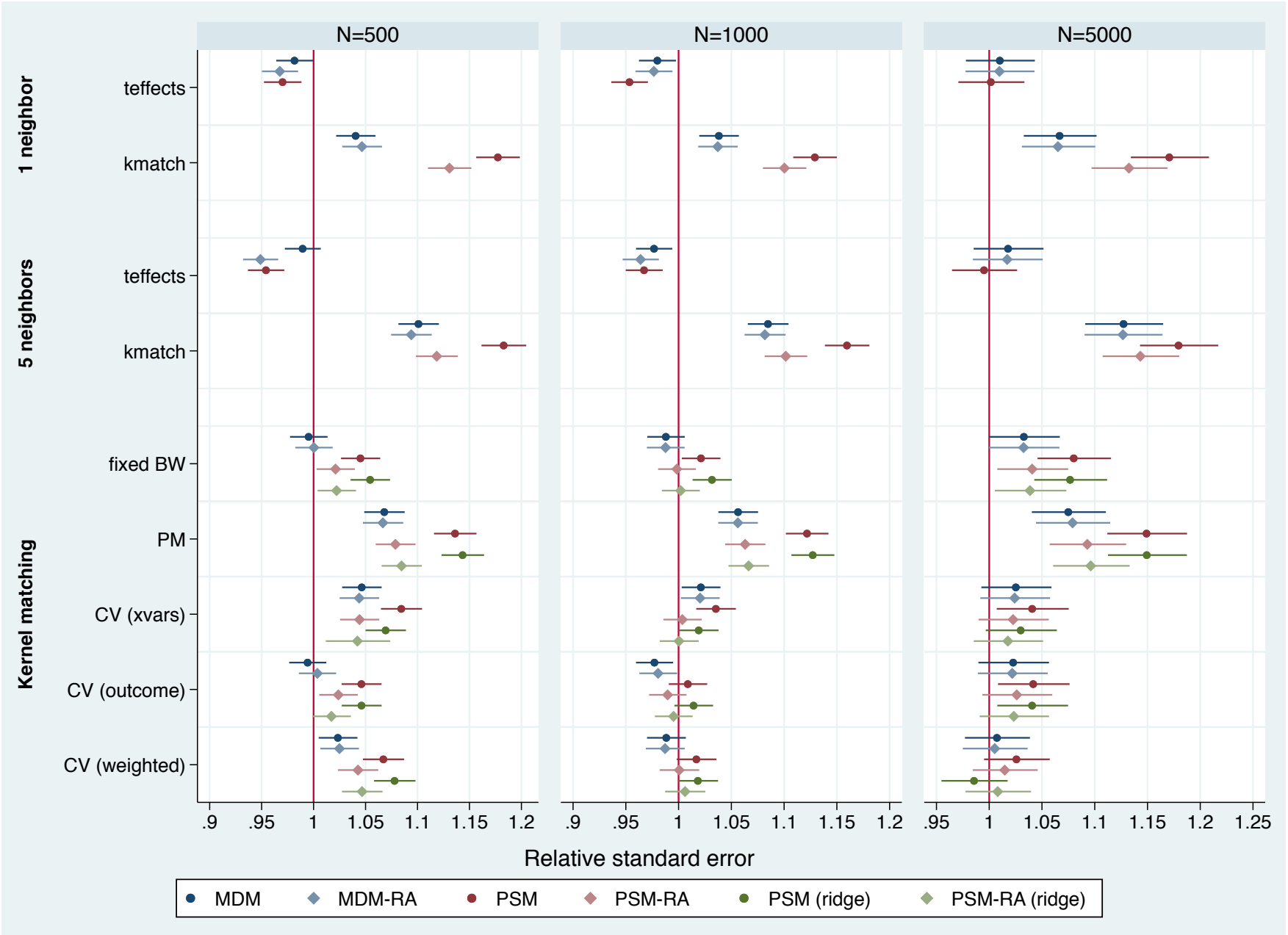
Results: Mean Squared Error



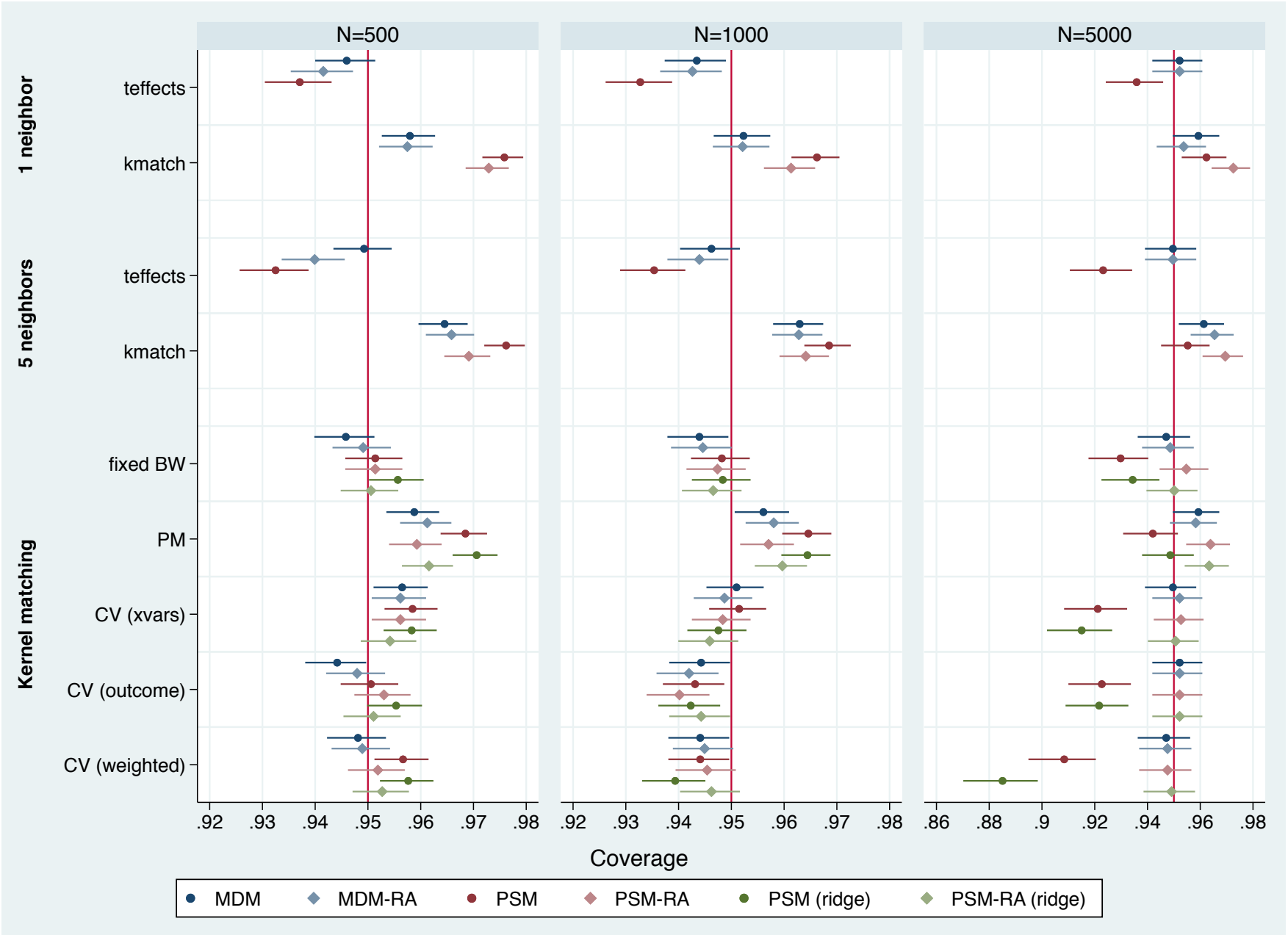
Results: Bias Reduction



Results: Validity of Bootstrap Standard Errors



Results: Validity of Bootstrap Confidence Intervals



1 Potential Outcomes and Causal Inference

2 Matching

3 Propensity Score Matching

4 King and Nielsen's "Why Propensity Scores Should Not Be Used for Matching"

5 Are King and Nielsen right?

6 Illustration using `kmatch`

7 Conclusions

Conclusions

- The arguments brought forward by King and Nielsen against Propensity Score Matching are valid, but they mostly apply to one specific form of PSM: one-to-one matching without replacement (pair matching).
- Other PSM matching algorithms perform much better because they are not affected by the random pruning problem.
- Theoretical results (see, e.g., Frölich 2007) suggest, that MDM will tend to outperform PSM in terms of efficiency also for these algorithms, but the differences are likely to be small.

Conclusions

- Some conclusions from the simulation
  - For PSM, application of regression-adjustment seems like a great idea (reduction of bias and variance); for MDM the advantages regression-adjustment are less clear.
  - Bootstrap standard error/confidence interval estimation seems to be mostly ok for kernel/ridge matching; this is in contrast to nearest-neighbor matching, where bootstrap standard errors are clearly biased.
- To do
  - Run some simulations comparable to the ones by King and Nielsen using different algorithms.

References I

- Blackwell, M., S. Iacus, G. King, Giuseppe Porro. 2009. cem: Coarsened exact matching in Stata. *The Stata Journal* 9(4): 524–546.
- Cochran, W.G. 1968. The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies. *Biometrics* 24(2):295–313.
- Frölich, M. 2007. On the inefficiency of propensity score matching *AStA* 91:279–290.
- Hendrickx, J. 2002. ISKO: Stata module to recode 4 digit ISCO-88 occupational codes. Statistical Software Components S425802, Boston College Department of Economics.
- Holland, Paul W. (1986). Statistics and Causal Inference. Journal of the American Statistical Association 81: 945-960.
- King, G., R. Nielsen. 2016. Why Propensity Scores Should Not Be Used for Matching. Working Paper. Available from http://j.mp/1sexgVw.

References II

- Mill, J.S. 2002. *A System of Logic*. Reprinted from the 1981 edition (first published 1843). Honolulu, Hawaii: University Press of the Pacific.
- Neyman, J. 1990[1923]. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9 (Translated and edited by D.M. Dabrowska and T.P. Speed from the Polish original). *Statistical Science* 5(4):465–472.
- Rosenbaum, P.R., D.B. Rubin. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 70:41–55.
- Rubin, D.B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5):688–701.
- Rubin, D.B. 1990. Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies. *Statistical Science* 5(4):472–480.